Towards FDG-PET image characterization and classification : application to Alzheimer's disease computer-aided diagnosis

Xiaoxi Pan

École Doctorale : Ecole Doctorale Physique et Sciences de la Matière (ED352)

Laboratoire de l'Institut Fresnel

THÈSE DE DOCTORAT

pour obtenir le grade de

DOCTEUR de l'ÉCOLE CENTRALE de MARSEILLE

Discipline : Optique, Photonique et Traitement d'Image

Towards FDG-PET image Characterization and Classification: Application to Alzheimer's Disease Computer-aided Diagnosis

par

Xiaoxi PAN

Directeur de thèse : Prof. Mouloud ADEL

Co-directrice de thèse : Prof. Caroline FOSSATI

Soutenue le 28 octobre 2019

devant le jury composé de :

Prof. Régine LE BOUQUIN JEANNES	Université de Rennes 1	Rapporteur
Prof. Amine NAIT-ALI	Université Paris-Est Créteil	Rapporteur
Prof. Didier WOLF	Université de Lorraine	Examinateur
Prof. Eric GUEDJ	Aix-Marseille Université	Examinateur
Prof. Mouloud ADEL	Aix-Marseille Université	Directeur de thèse
Prof. Caroline FOSSATI	Ecole Centrale de Marseille	Co-directrice de thèse

Acknowledgments

I am thankful that Dr. Régine Le Bouquin Jeannès and Dr.Amine Nait-Ali have accepted to be the reviewers of my thesis. Thanks to Dr. Didier Wolf and Dr. Eric Guedj for agreeing to be the examiners.

I would like to express my deepest appreciation to my supervisors, Mouloud Adel and Caroline Fossati, for receiving me to start my PhD journey in France. Thanks for their guidance, support, and encouragement, spreading from work to life. Thanks for their patience and tolerance for troubles that I may cause. Thank Thierry Gaidon for his suggestions and opinions shared in each meeting. Thank Julien Wojak for his technical support, especially for his patient answers to my simple questions. Thank Eric Guedj for his guidance and valuable comments from medical aspects. Moreover, thank Salah Bourennane for his kind help and care in the office and other team members, in spite of a short time spent with them, I am still happy to know them.

I also would like to thank my super lovely and kind friends that I have met in Marseille, it was they who accompanied me for more than a thousand days and because of them, I have never felt lonely and helpless. I am very happy to have them during my PhD study, they are Lu Ren, Jie Zhang, Dan Feng, Lingyu Kong, Guochao Gao, Qi Wang, Qiaoqiao Sun, Jian Yang, Jinming Lv, Wenjing Yang, Lan Yang, and a special friend, Ruotian Liu. Thanks to him for being in my life finally.

Furthermore, thanks to my grandparents who have accompanied me for my childhood and thanks to my parents and my brother for their love, understanding, and support. I love them from my heart as always happens.

Last but not least, thanks to China Scholarship Council (CSC) for their funding support during the three years.

Publications

- Journal papers

    1. **X. Pan**, M. Adel, C. Fossati, T. Gaidon, J. Wojak and Eric Guedj. Multi-scale Spatial Gradient Features for 18F-FDG PET Image-Guided Diagnosis of Alzheimer's Disease. *Computer Methods and Programs in Biomedicine*, vol. 180, 2019.
    2. **X. Pan**, M. Adel, C. Fossati, T. Gaidon, and Eric Guedj. Multilevel Feature Representation of FDG-PET Brain Images for Diagnosing Alzheimer's Disease. *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1499-1506, 2019.

- Conference papers

    1. **X. Pan**, M. Adel, C. Fossati, T. Gaidon, J. Wojak and Eric Guedj. First and Second Order Gradients for Alzheimer's Disease Diagnosis. *5th International Conference on Frontiers of Signal Processing (ICFSP)*. (Excellent oral presentation)
    2. **X. Pan**, M. Adel, C. Fossati, T. Gaidon, and Eric Guedj. Alzheimer'S Disease Diagnosis with FDG-PET Brain Images By Using Multi-Level Features. *25th IEEE International Conference on Image Processing (ICIP)*. (Oral)

- Book chapter

    1. M. Adel, I. Garali, **X. Pan**, C. Fossati, T. Gaidon, J. Wojak, S. Bourennane and E. Guedj. Alzheimer's Disease Computer-Aided Diagnosis on Positron Emission Tomography Brain Images Using Image Processing Techniques. *Computer Methods and Programs in Biomedical Signal and Image Processing*, DOI: 10.5772/intechopen.86114.

Contents

List of Figures **ix**

List of Tables **xiii**

Abstract **1**

Résumé **3**

Présentation du travail en Français **5**

1 Introduction **11**

1.1 Alzheimer's disease 11

1.2 Neuroimaging modalities 14

1.2.1 Magnetic resonance imaging 14

1.2.2 Functional magnetic resonance imaging 14

1.2.3 Diffusion tensor imaging 15

1.2.4 Single photon emission computed tomography 16

1.2.5 Positron emission tomography 16

1.3 Contributions and outline 19

2 Computer-aided Diagnosis Methods: Application to Alzheimer's Disease **23**

2.1 Introduction 23

2.2 Discriminative features for representations 24

2.2.1 Voxel-wise features 24

2.2.2 ROI-wise features 24

2.2.3 Feature selection/reduction 26

2.3 Machine learning techniques for classifications 29

2.3.1 Logistic Regression 29

2.3.2 Support Vector Machine . . . 32
2.3.3 Ensemble methods . . . 35
2.4 Performance evaluation . . . 37
2.4.1 Evaluation measures . . . 38
2.4.2 Cross-validation . . . 39
2.5 Application to Alzheimer's disease diagnosis . . . 41
2.6 Deep learning techniques . . . 41
2.6.1 MultiLayer Perceptron . . . 43
2.6.2 Convolutional Neural Network . . . 44
2.6.3 Applications to neuroimaging . . . 46
2.7 Conclusion . . . 47

3 ADNI Data **49**
3.1 Introduction . . . 49
3.2 Data acquisition . . . 49
3.2.1 FDG-PET in ADNI . . . 49
3.2.2 FDG-PET downloaded from IDA . . . 51
3.3 Data selection . . . 54
3.4 Data processing . . . 57
3.5 Conclusion . . . 58

4 Multilevel Feature Representation for FDG-PET Images **61**
4.1 Introduction . . . 61
4.2 Method . . . 61
4.2.1 Feature extraction . . . 62
4.2.2 Feature selection . . . 70
4.2.3 Ensemble classification . . . 70
4.3 Experiments and results . . . 71
4.3.1 Setup . . . 71
4.3.2 Single-type feature representation evaluation . . . 72
4.3.3 Feature selection evaluation . . . 73
4.3.4 Feature concatenation evaluation . . . 75
4.3.5 Effectiveness of the similarity-driven ranking method . . . 76
4.3.6 Ensemble classification evaluation . . . 77
4.4 Conclusion . . . 78

5 Multiscale Spatial Gradient Features for Characterizing FDG-PET Images **81**
5.1 Introduction . . . 81
5.2 Method . . . 81
5.2.1 Histogram of Oriented Gradient for 2D images . . . 82
5.2.2 Histogram of Oriented Gradient for FDG-PET images . . . 82
5.2.3 ROI ranking . . . 86
5.2.4 Ensemble classification . . . 86
5.3 Experiments and results . . . 88
5.3.1 Setup . . . 88
5.3.2 Evaluation on spatial gradient features for FDG-PET images . 89
5.3.3 Evaluation on ROI ranking method . . . 90
5.3.4 Evaluation on ensemble classification . . . 91
5.4 Conclusion . . . 94

6 Multiview Convolutional Neural Network for AD diagnosis and MCI Conversion Prediction **95**
6.1 Introduction . . . 95
6.2 Method . . . 95
6.2.1 Multiview CNN architecture . . . 97
6.2.2 Implementations . . . 99
6.3 Experiments and results . . . 101
6.3.1 Setup . . . 101
6.3.2 Evaluation on single view CNN . . . 102
6.3.3 Evaluation on multiple view CNN . . . 104
6.3.4 Comparison with state-of-the-art methods . . . 105
6.4 Conclusion . . . 107

7 Conclusions and Perspectives **109**

Bibliography **111**

List of Figures

1.1 The cerebral cortex. 12
1.2 The timeline of AD progression. 12
1.3 Biomarker changes in the progression of AD. 13
1.4 Differences between NC and AD captured by MRI scan (coronal view), from Alzheimer's Disease Neuroimaging Initiative (ADNI) . . . 14
1.5 Differences of functional connectivity between NC and AD captured by rs-fMRI (axial view). 15
1.6 Differences in FA (fractional anisotropy) image between NC and AD captured by DTI (coronal view), from ADNI. 16
1.7 Differences between NC and AD captured by SPECT (axial view), from ADNI. 17
1.8 Differences between NC and AD captured by PiB-PET (axial view), from ADNI. 18
1.9 Differences between NC and AD captured by Florbetapir-PET (axial view), from ADNI. 18
1.10 Differences between NC and AD captured by FDG-PET (axial view), from ADNI. 19

2.1 CAD framework for Alzheimer's disease. 23
2.2 Logistic function. 30
2.3 Maximum-margin hyperplane and margins for an SVM. 33
2.4 An ensemble architecture. 35
2.5 An instance of ROC curve. 39
2.6 An instance of cross-validation. 40
2.7 An instance of MultiLayer Perceptron 44
2.8 An illustration of Relu and tanh functions 44
2.9 An illustration of convolution and pooling 45
2.10 An illustration of general CNN architecture. 46

3.1 Image Searching Options 51
3.2 Different image types involved in ADNI. 54
3.3 An instance of different image types. (a) Original FDG-PET data. (b) Pre-processed data. (c) Post-processed data 55
3.4 FDG-PET Image Searching Options 56
3.5 FDG-PET Image Processing Pipeline 58

4.1 The framework of the proposed method. 62
4.2 Statistics of the similarity coefficients between subjects for a certain ROI. Top row: AD vs. NC. Bottom row: pMCI vs. sMCI. 65
4.3 Instance of the division for a similarity matrix $\mathbf{W_r}$. 66
4.4 Instance of the brain connectivity network from the axial view. 66
4.5 Performance of feature selection under different values of λ and a smaller λ implies less features are selected. (a) AD vs. NC. (b) pMCI vs. sMCI. 75
4.6 Performance evaluation of the similarity-driven ranking method. (a) AD vs. NC. (b) pMCI vs. sMCI. 77
4.7 Performance evaluation of the ensemble classification. (a) AD vs. NC. (b) pMCI vs. sMCI. 78

5.1 The flowchart of the proposed method. 82
5.2 An instance to differentiate between NC and AD via intensity and gradient, where the top row is for an NC subject and the bottom row is for a subject with AD. (a) A slice of a subject. (b) Part of region Parietal_Inf_R. (c) The corresponding gradient. 83
5.3 An instance of the polar angle ϕ and azimuth angle θ. 84
5.4 Histogram of Oriented Gradient (HOG) for an NC subject (top row) and an AD subject (bottom row) without segmentation. (a) 18×18 ($\phi \times \theta$) bins. (b) 18×9 bins. (c) 12×6 bins. 85
5.5 The framework of region ranking method. 86
5.6 The framework of the ensemble classification, circles with different colors indicate different regions. 88
5.7 Maximum difference of accuracy (or balanced accuracy) (a), and area under ROC (b), in five scales of SSH for each ROI 93
5.8 Performance under different numbers of ROIs. (a) AD vs. NC. (b) pMCI vs. sMCI. 94

6.1 The proposed multiview CNN architectures. (a) mvCNNiF. (b) mvCNNaF. 96
6.2 Cuboïd kernel for each view. 97
6.3 Details of blocks. (a) ConvBlock. (b) FCBlock. 98
6.4 The proposed multiview CNN architectures. (a) mvCNNiF. (b) mvCNNaF. 100
6.5 An instance of the mapping procedure. 100
6.6 Training and validation losses of different views during a training procedure where the top row is for AD vs.NC and the bottom row is for pMCI vs. sMCI. (a) Axial view. (b) Coronal view. (c) Sagittal view. 104
6.7 Performance of multiview CNN. (a) AD vs. NC. (b) pMCI vs. sMCI. 105
6.8 Training and validation losses of mvCNNiF. (a) AD vs. NC (b) pMCI vs. sMCI. 105

List of Tables

1.1 Neuroimaging modalities for AD diagnosis. 19

2.1 A brief overview of features used in CAD methods for Alzheimer's disease. 28

2.2 A confusion matrix for a binary classification. 38

2.3 A brief overview of classifiers used in CAD methods for Alzheimer's disease. 42

2.4 A brief overview of deep learning techniques used in CAD methods for Alzheimer's disease. 48

3.1 Demographic and clinical information of subjects. 59

4.1 Top 20 ROIs highly relevant to AD or pMCI 67

4.2 Top 20 ROIs slightly relevant to AD or pMCI, Rank -1 indicates the most irrelevant to AD or pMCI 68

4.3 Performance of different types of feature for AD vs. NC(%) 73

4.4 Performance of different types of feature for pMCI vs. sMCI(%) . . . 73

4.5 Performance of feature selection for AD vs. NC(%) 74

4.6 Performance of feature selection for pMCI vs. sMCI(%) 74

4.7 Performance of different levels of feature for AD vs. NC(%) 76

4.8 Performance of different levels of feature for pMCI vs. sMCI(%) . . . 76

5.1 Comparison of common features and spatial gradient features for AD vs. NC(%) 89

5.2 Comparison of common features and spatial gradient features for pMCI vs. sMCI(%) 89

5.3 Top 20 ROIs for AD vs. NC and pMCI vs. sMCI 92

6.1 mvCNNaF architecture hyparameters and output size for **axial and sagittal** views 101
6.2 mvCNNaF architecture hyparameters and output size for **coronal** view 102
6.3 Performance of single view for AD vs. NC (%) 103
6.4 Performance of single view for pMCI vs. sMCI (%) 103
6.5 Performance comparison with state-of-the-art methods for AD vs. NC(%) 106
6.6 Performance comparison with state-of-the-art methods for pMCI vs. sMCI(%) 106

Abstract

Alzheimer's disease (AD) is becoming the dominant type of neurodegenerative brain disease in elderly people, which is incurable and irreversible for now. It is expected to diagnose its early stage, Mild Cognitive Impairment (MCI), then interventions can be applied to delay the onset. Fluorodeoxyglucose positron emission tomography (FDG-PET) is considered as a significant and effective modality to diagnose AD and the corresponding early phase since it can capture metabolic changes in the brain thereby indicating abnormal regions. Therefore, this thesis is devoted to identifying AD from Normal Control (NC) and predicting MCI conversion under FDG-PET modality. For this purpose, three independent novel methods are proposed.

The first method focuses on developing connectivities among anatomical regions involved in FDG-PET images which are rarely addressed in previous methods. Such connectivities are represented by either similarities or graph measures among regions. Then combined with each region's properties, these features are fed into a designed ensemble classification framework to tackle problems of AD diagnosis and MCI conversion prediction.

The second method investigates features to characterize FDG-PET images from the view of spatial gradients, which can link the commonly used features, voxel-wise and region-wise features. The spatial gradient is quantified by a 2D histogram of orientation and expressed in a multiscale manner. The results are given by integrating different scales of spatial gradients within different regions.

The third method applies Convolutional Neural Network (CNN) techniques to three views of FDG-PET data, thereby proposing the main multiview CNN architecture. Such an architecture can facilitate convolutional operations, from 3D to 2D, and meanwhile consider spatial relations, which is benefited from a novel mapping layer with cuboid convolution kernels. Then three views are combined and make a decision jointly.

Experiments conducted on public dataset show that the three proposed methods

can achieve significant performance and moreover, outperform most state-of-the-art approaches.

Keywords: Feature extraction, Classification, Convolutional Neural Network, Computer-aided diagnosis, FDG-PET, Alzheimer's disease

Résumé

La maladie d'Alzheimer (MA) est la maladie neurodégénérative - incurable et irréversible pour le moment - la plus répandue chez les personnes âgées. On s'attend à ce qu'elle soit diagnostiquée à son stade précoce, Mild Cognitive Impairment (MCI), pour pouvoir intervenir et retarder son apparition. La tomographie par émission de positons au fluorodésoxyglucose (TEP-FDG) est considérée comme une modalité efficace pour diagnostiquer la MA et la phase précoce correspondante, car elle peut capturer les changements métaboliques dans le cerveau, indiquant ainsi des régions anormales. Cette thèse est consacrée à identifier et distinguer, sur des images TEP, les sujets atteints de MA de ceux qui sont sains. Ce travail vise également à prédire la conversion de MCI sous la modalité d'imagerie TEP-FDG. A cette fin, trois nouvelles méthodes indépendantes sont proposées.

La première méthode est axée sur le développement de connectivités entre les régions anatomiques impliquées dans les images au TEP-FDG, qui sont rarement abordées dans les méthodes déjà publiées. Ces connectivités sont représentées par des similarités ou des mesures graphiques entre régions. Combinées ensuite aux propriétés de chaque région, ces caractéristiques sont intégrées dans un cadre de classification d'ensemble conçu pour résoudre les problèmes de diagnostic MA et de prédiction de conversion MCI.

La seconde méthode étudie les caractéristiques permettant de caractériser les images au TEP-FDG à partir de gradients spatiaux, ce qui permet de lier les caractéristiques couramment utilisées, voxel ou régionales. Le gradient spatial est quantifié par un histogramme 2D d'orientation et exprimé sous forme multi-échelle. Les résultats sont obtenus en intégrant différentes échelles de gradients spatiaux dans différentes régions.

La troisième méthode applique le réseau neuronal convolutif sur les trois axes des données 3D de TEP-FDG, proposant ainsi la principale architecture CNN à vues multiples. Une telle architecture peut faciliter les opérations de convolution, de la 3D à la 2D, tout en tenant compte des relations spatiales, qui bénéficient d'une

nouvelle couche de cartographie. Ensuite, les traitements sur les trois axes sont combinés et prennent une décision conjointement.

Les expériences menées sur des ensembles de données publics montrent que les trois méthodes proposées peuvent atteindre des performances significatives et, de surcroît, dépasser les approches les plus avancées.

Mots clés: Extraction de caractéristiques, Classification, Réseau neuronal convolutif, Diagnostic assisté par ordinateur, TEP-FDG, La maladie d'Alzheimer

Présentation du travail en Français

La maladie d'Alzheimer (MA), décrite pour la première fois par Alois Alzheimer, psychiatre et pathologiste allemand, est une maladie neurodégénérative irréversible et le type de démence le plus répandu. On estime que près de 70% des facteurs de risque sont liés à l'hérédité et que d'autres facteurs de risque incluent des antécédents de traumatisme crânien, de dépression et d'hypertension. La progression de la maladie est liée à l'accumulation de plaques amyloïdes (Aβ et tau) et d'enchevêtrements neurofibrillaires dans le cerveau.

La maladie d'Alzheimer survient généralement dans les lobes temporaux et pariétaux associés à la mémoire et au langage. Le symptôme précoce le plus courant est la perte de mémoire à court terme. Différents symptômes peuvent apparaître progressivement au fur et à mesure de l'évolution de la maladie: troubles du langage, désorientation, instabilité émotionnelle, perte de motivation et nombreux problèmes de comportement. Avec l'aggravation de la situation, les patients ont tendance à perdre progressivement leurs fonctions physiques, entraînant éventuellement la mort. La maladie d'Alzheimer affecte une personne sur neuf âgée de plus de 65 ans et une sur trois âgée de plus de 85 ans. On estime que 131 millions de personnes vivront avec la maladie en 2050. A l'heure actuelle, la maladie d'Alzheimer reste incurable, même si des développements prometteurs en matière de traitement sont espérés dans un proche avenir.

La tomographie par émission de positrons (TEP) est une méthode d'imagerie fonctionnelle en médecine nucléaire qui utilise un radiotraceur dont l'activité est détectée dans le corps pour obtenir des informations sur l'activité cellulaire ou des informations métaboliques facilitant ainsi le diagnostic médical. Il existe différents types de radiotraceurs qui peuvent être utilisés en TEP, tels que le fluorodésoxyglucose (^{18}F-FDG), le composé B de ^{11}C-Pittsburgh (^{11}C-PiB) et le florbétapir ^{18}F-florbetapir (^{18}F-AV-45), *etc.* La TEP-^{18}F-FDG (appelée ci-après TEP-FDG) est utilisée pour mesurer l'absorption de glucose par les neurones et les cellules gliales qui est considérée comme un indicateur sensible des modifications de la fonction synap-

tique. Les patients atteints de la maladie d'Alzheimer se révèlent avoir de graves anomalies du métabolisme du glucose. Par conséquent, les zones particulièrement affectées par la maladie peuvent être détectées ou localisées via le radiotraceur, le ^{18}F-FDG, qui reflète le métabolisme du glucose. Un certain nombre d'études sur la TEP-FDG ont montré que les patients atteints de MA présentaient un métabolisme localement plus faible et significatif, y compris au niveau du lobe temporal, du lobe pariétal, du gyrus cingulaire postérieur avec une expansion vers le lobe frontal lorsque la maladie se développait, par rapport au groupe normal, alors que le gyrus central antérieur et postérieur, le cervelet et le thalamus sont relativement normaux. De plus, on pense que des modifications de l'activité métabolique sont nécessaires avant l'atrophie de la structure. Par conséquent, la TEP-FDG est généralement considérée comme l'une des modalités les plus efficaces pour un diagnostic précoce.

Cette thèse est consacrée à l'exploitation d'images TEP-FDG combinées à des méthodes d'apprentissage automatique pour diagnostiquer la maladie d'Alzheimer et prédire la conversion de sa phase initiale, Mild Cognitive Impairment (MCI) en une phase de maladie avérée. De telles méthodes visent à aider les médecins à effectuer une analyse et une évaluation complètes dans un court laps de temps, à identifier les zones présentant des risques potentiels et à donner ainsi des suggestions de référence. En outre, la prévision de la progression du MCI peut également éviter efficacement les risques et retarder l'apparition de la maladie. Cette thèse réalisée dans cette optique comporte 7 chapitres.

Chapitre 1. Introduction

Ce chapitre porte principalement sur la maladie d'Alzheimer, ses symptômes, sa progression et ses effets sur l'homme, en particulier chez les personnes âgées. Ensuite, différentes méthodes et indicateurs de diagnostic clinique sont présentés. Enfin, nous comparons diverses modalités de neuro-imagerie pour le diagnostic de la MA, notamment l'imagerie par résonance magnétique (IRM), l'IRM fonctionnelle (IRMf), l'imagerie par tenseur de diffusion (ITD), la tomographie par emission mono-photonique (TEMP) et la tomographie par émission de positrons (TEP). A la fin du chapitre, l'esquisse de la thèse, ainsi que ses principales contributions, sont présentées.

Chapitre 2. Méthodes de diagnostic assisté par ordinateur: application à la maladie d'Alzheimer

Dans ce chapitre, deux parties relevant des méthodes de diagnostic assisté par ordinateur sont présentées, comprenant l'extraction et la classification de caractéristiques. L'extraction de caractéristiques est abordée dans un premier temps, il y est question principalement des caractéristiques couramment utilisées et de nouvelles, telles que l'intensité moyenne régionale, l'écart-type, les caractéristiques textures, etc. Les techniques de réduction ou de sélection des caractéristiques correspondantes sont ensuite présentées, par exemple les méthodes de filtre, méthodes d'enveloppe et méthodes d'intégration, suivis d'une brève revue des méthodes d'extraction de caractéristiques dans la littérature récente rattachée au domaine de la thèse. La deuxième partie de ce chapitre donne un aperçu des concepts de machine learning pour la classification, y compris des classificateurs, tels que Régression logistique, Machine à vecteurs de support, et les mesures d'évaluation permettant d'évaluer les performances de la classification, telles que la précision et la sensibilité, spécificité et aire sous la courbe. En outre, des modèles d'apprentissage profond (deep learning) sont également décrits, y compris Perceptron multicouche et Réseau neuronal convolutif, suivis d'applications au diagnostic de la maladie d'Alzheimer.

Chapitre 3. Données ADNI

La plupart des méthodes de classification de la MA en imagerie neurologique sont généralement testées sur un ensemble de données public, les données ADNI (Alzheimer's Disease Neuroimaging Initiative). Cependant, en raison de la complexité et de la diversité des données, la plupart des méthodes n'ont exploité qu'un sous-ensemble et peu d'entre elles ont clarifié les détails de la sélection des données. Au chapitre 3, nous clarifions étape par étape la procédure d'acquisition de données dans la base ADNI et la règle de sélection des données afin de permettre une comparaison équitable avec d'autres méthodes. De plus, nous avons obtenu un total de 1048 images TEP-FDG à partir des balayages de base de 1048 sujets, respectivement. Il s'agit en fait de l'ensemble de données le plus complet permettant d'assurer l'efficacité des évaluations des méthodes proposées.

Chapitre 4. Représentation de caractéristique multiniveau pour les images TEP-FDG

Au chapitre 4, la représentation de caractéristiques multiniveaux pour les données TEP-FDG est étudiée pour diagnostiquer la maladie d'Alzheimer et son stade précoce. Premièrement, après avoir segmenté chaque sujet en 90 régions selon un atlas AAL (Automated Anatomical Labeling), 3 niveaux de caractéristiques sont extraits, plus précisément les caractéristiques de niveau 1, qui comprennent l'intensité moyenne et l'écart-type de la région. La caractéristique de deuxième niveau, la connectivité basée sur la similarité entre n'importe quelle paire de régions, est décomposée en 3 ensembles selon une méthode de classement que nous proposons et qui est basée sur la similarité. La caractéristique de troisième niveau est composée de mesures issues des graphes. Ensuite, une stratégie de sélection d'options, Least Absolute Shrinkage and Selection Operator (LASSO), est appliquée à chaque ensemble de caractéristiques. Différents classifieurs sont alors construits à partir de différents ensembles de caractéristiques. La prédiction finale est obtenue par un classifieur d'ensemble dont le choix est fait par une stratégie proposée associée à une technique de validation croisée imbriquée. Les principales contributions obtenues dans ce chapitre peuvent être résumées en trois volets: 1) la représentation des entités à plusieurs niveaux prend en compte non seulement les propriétés de région, mais également la connectivité entre deux paires de régions et une connectivité globale entre une région et les autres; 2) une méthode de classement basée sur la similarité est proposée pour classer les régions des plus affectées aux moins affectées par la maladie, ce qui peut réduire la dimension et augmenter dans une certaine mesure la diversité du classifieur; 3) une stratégie de sélection de classifieur est proposée pour choisir une paire de classifieurs avec une grande diversité afin d'améliorer l'effet d'ensemble, en particulier dans le cas où les sous-classifieurs ne conduiraient pas a des résultats suffisants.

Chapitre 5. Caractéristiques de gradients spatiaux multiéchelles pour la caractérisation d'images TEP-FDG

Dans ce chapitre, les caractéristiques utilisées pour caractériser les images au TEP-FDG sont extraites d'un autre point de vue: les gradients spatiaux des taux

de FDG dans les images du cerveau en TEP, au lieu de caractéristiques par voxels et par régions, comme de nombreuses études l'ont fait auparavant. Ce travail est motivé par les différences observées des taux de FDG entre sujets AD et Normal Control (NC). Les gradients spatiaux sont quantifiés par un histogramme d'orientation 2D, similaire à l'Histogram of Oriented Gradient (HOG) appliqué avec succès à la détection d'objets dans des images 2D. Tout d'abord, le gradient spatial de l'image au TEP-FDG est calculé, puis 90 régions sont extraites de l'image du gradient par le biais d'un atlas AAL, dans lequel le cervelet n'est pas pris en compte. Ensuite, certaines régions distinctes sont sélectionnées via une méthode de classement de régions que nous avons proposée et qui prend en compte plusieurs descripteurs Small Scale HOG (SSH) de chaque région. Enfin, un classificateur d'ensemble est formé dans les régions sélectionnées à l'aide des fonctions SSH et LSH (Large Scale HOG). Les contributions impliquées dans ce chapitre peuvent être résumées en trois aspects: 1) Le descripteur 1D HOG, utilisé à l'origine dans les images de scènes naturelles, est amélioré en 2D HOG pour quantifier les gradients spatiaux, caractérisant ainsi les images cérébrales 3D TEP-FDG. De plus, 2D HOG est exprimé en SSH et LSH, ce qui s'avère plus efficace que les caractéristiques couramment utilisées; 2) une méthode de classement de régions est proposée pour sélectionner des régions distinctes en utilisant plusieurs caractéristiques SSH; 3) un cadre de classification d'ensemble est conçu en prenant en compte les HOG 2D en 2D à petite et grande échelle pour la région individuelle et les régions concaténées, ce qui améliore la précision de la classification pour le diagnostic.

Chapitre 6. Convolutional Neural Network multi-vues pour le diagnostic de la maladie d'Alzheimer et la prévision de conversion de MCI

Compte tenu des performances impressionnantes obtenues grâce au deep learning, au chapitre 6, nous essayons de diagnostiquer la MA et de prédire la conversion du MCI dans le cadre de CNN. En conséquence, deux architectures CNN à vues multiples sont proposées, notées mvCNNiF et mvCNNaF, avec des différences dans la combinaison de vues multiples, y compris des vues axiales, coronales et sagittales. mvCNNiF consiste à concaténer plusieurs vues au niveau de la première couche entièrement connectée de chaque branche, puis à passer trois autres couche entièrement connectées avant de prendre une décision. Par conséquent, trois vues

impliquées dans l'architecture sont traitées simultanément. mvCNNaF consiste à intégrer les résultats de plusieurs vues après la dernière couche entièrement connectée de chaque branche par un vote à la majorité. Avec cette architecture, les modèles des vues axiales, coronales et sagittales sont formés séparément la décision est prise ensuite conjointement. En dehors de cela, une autre contribution principale de ce chapitre réside dans une couche de cartographie avec des noyaux de convolution cuboïdes conçue pour projeter des informations le long de la troisième dimension sur un plan. La couche de cartographie proposée peut réduire les paramètres impliqués dans les couches de convolution suivantes et, dans l'intervalle, envisager des relations spatiales dans différentes vues. A la fin du chapitre 6, nous comparons les trois méthodes proposées avec des algorithmes à la pointe de la technologie. Les résultats indiquent que les trois méthodes proposées affichent des performances de pointe en diagnostic de la MA, alors que pour la prévision de conversion MCI, ces méthodes fonctionnent bien mais pourraient être encore améliorées.

Chapitre 7. Conclusion et perspectives

Ce chapitre conclut d'abord le travail de la thèse, met en évidence les contributions et souligne la nécessité d'améliorations. Ensuite, les travaux futurs sont introduits à partir des trois aspects, données, méthodes et tâches. Pour les données, plusieurs modalités, telles que la combinaison de l'IRM et de la TEP-FDG, peuvent être utilisées pour résoudre le problème du diagnostic de la MA. En outre, les données longitudinales constituent également un sujet important, qui peut non seulement élargir l'ensemble de données, mais également fournir un aperçu de la progression de la MA. Pour les méthodes, les déviations de second ordre, LBP (Local Binary Pattern) ou SIFT (Scale Invariant Feature Transform) méritent d'être étudiées. En outre, le réseaux antagonistes génératifs peut également être appliqué pour apprendre automatiquement des fonctions utiles. Pour les tâches, la prévision de mesures cliniques telles que le mini-examen de l'état mental et la localisation de régions à faible métabolisme sont des sujets importants pour expliquer et comprendre la maladie d'Alzheimer.

Chapter 1

Introduction

1.1 Alzheimer's disease

Alzheimer's disease (AD), first described by and named after German psychiatrist and pathologist Alois Alzheimer [1], is an irreversible neurodegenerative disease and the most common type of dementia. It is believed that nearly 70% of risk factors are related to heredity, and other risk factors include a history of head injuries, depression, and hypertension. The progression of the disease is related to the accumulation of plaques (Aβ and tau) and neurofibrillary tangles in the brain [2].

Alzheimer's disease usually occurs in temporal and parietal lobes which are associated with memory and language, as shown in Figure 1.1, an illustration of the cerebral cortex[1]. The most common early symptom is loss of short-term memory, and as the disease progresses, different symptoms may gradually appear, including language disorders, disorientation, emotional instability, loss of motivation and many behavioral problems [3, 4]. With the situation worsening, patients tend to gradually lose their physical function, eventually leading to death [3]. Alzheimer's disease affects one in nine over 65s [5] and one in three over 85s [6]. It is estimated that 131 million people will be living with the disease in 2050 [7]. At present, AD is incurable, even if promising developments for treatments are expected to be achieved in the near future.

The National Institute on Aging and Alzheimer's Association (NIA-AA) distinguishes 3 clinical stages: asymptomatic pre-clinical phase (pre-clinical stage of AD), amnestic Mild Cognitive Impairment (MCI) phase due to AD, and AD dementia phase [8–10], as illustrated in Figure 1.2 [8]. Pre-clinical AD is a stage in which the

[1] http://www.richardsonthebrain.com/cerebral-cortex

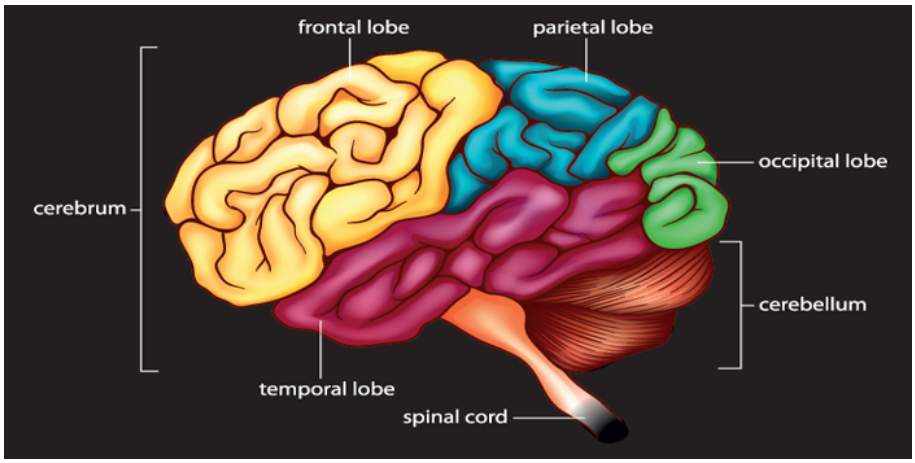


Figure 1.1: The cerebral cortex.

pathophysiological process of AD may begin many years before symptoms affecting memory, thinking or behavior can be detected [8]. People in the phase of amnestic MCI have more memory problems than normal aging people but do not yet meet the clinical criteria for AD. In addition, people with MCI have an increased risk of progressing to AD or another dementia. But it is worth noting that not all MCI will develop into dementia. In some cases, MCI reverts to normal cognition or remains stable. People with AD dementia are usually accompanied by significant symptoms, such as memory loss, word-finding difficulties, and visual/spatial problems, which are severe to impair a person's ability to live independently [3]. Considering AD is not curable, therefore any therapy or intervention for stages prior to AD would likely be of greatest benefit to delay the onset.

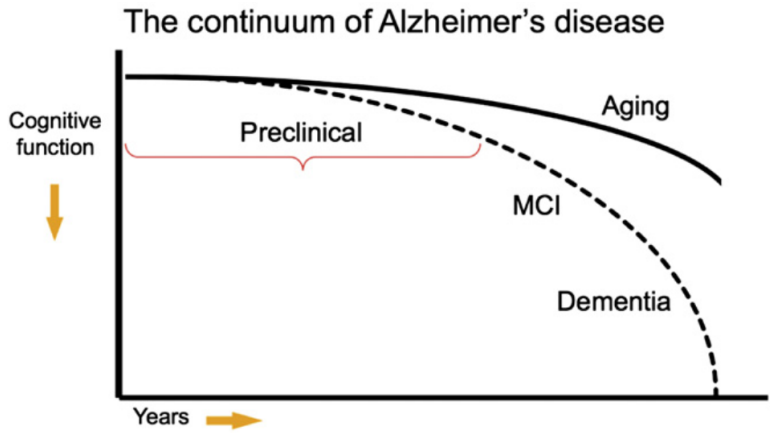


Figure 1.2: The timeline of AD progression.

The most commonly used criteria for diagnosis of Alzheimer's disease are provided by the National Institute of Neurological and Communicative Disorders and

Stroke - Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) [8–11]. Accordingly, AD is usually diagnosed based on the person's medical history, history from relatives, and clinical examination. Neuropsychological tests can further characterize the state of the disease, including a memory test such as California Verbal Learning Test (CVLT) [12], a language test such as Boston Naming Test (BNT) [13], or some comprehensive tests which combine a range of tests to provide an overview of cognitive skills, like mini–mental state examination (MMSE) and clinical dementia rating sum of boxes (CDR-SB) [14] *etc.* The NINCDS-ADRDA criterion has also suggested the utility of different biomarkers of the pathophysiological process to weight the diagnostic probability of the disease, such as cerebrospinal fluid (CSF) biomarkers (Aβ42, t-tau, p-tau), blood biomarkers, genetic biomarkers (APOE ε4), as well as neuroimaging modalities (PET, fMRI, SPECT). Figure 1.3 [8, 15] shows changes of different biomarkers during the progression to AD. As can be seen, biomarker changes have occurred before clinical dysfunction. Amyloid beta accumulation identified by CSF or PET is the most sensitive, followed by synaptic dysfunction. Early diagnosis becomes crucial to either allow patients to receive interventions at an early stage or provide insights into the disease progression. Therefore, the objective of this thesis is to exploit neuroimaging modalities combining with machine learning methods to identify patients with AD and patients with a high risk of cognitive decline from the Normal Control (NC).

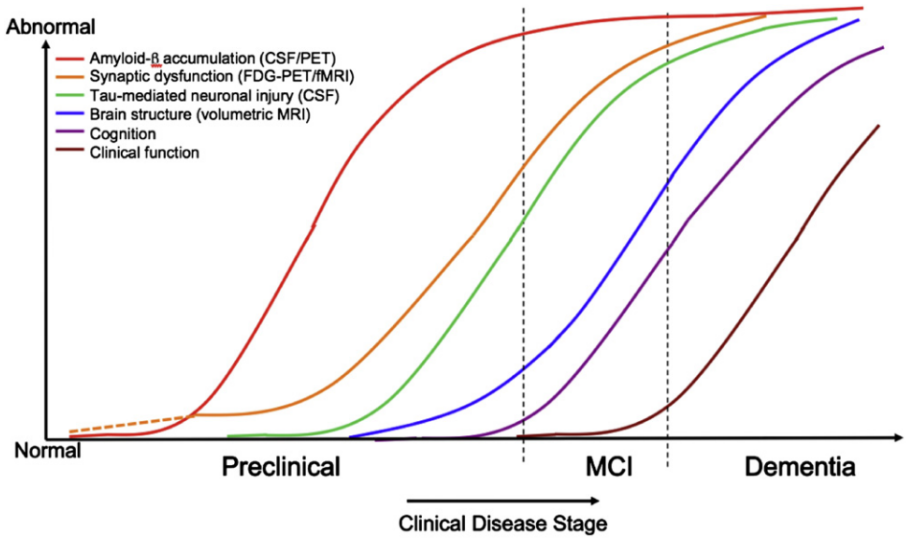


Figure 1.3: Biomarker changes in the progression of AD.

1.2 Neuroimaging modalities

With the rapid development of medical imaging, a variety of technologies have emerged, such as functional magnetic resonance imaging (fMRI), single photon emission computed tomography (SPECT), and positron emission computed tomography (PET), *etc.* They have their own characteristics, therefore have different applications, which provides a possibility for early diagnosis of AD.

1.2.1 Magnetic resonance imaging

Magnetic resonance imaging (MRI) is based on the theory of nuclear magnetic resonance to generate structural images of inner organs or tissues with high quality. MRI can display the brain anatomy with a high resolution and can clearly distinguish between gray matter and white matter. Since one of AD characters is the cortical shrinkage, structure MRI therefore provides guidelines for assessing brain atrophy in patients through measuring the regional or whole brain volume. Figure 1.4 shows the evident cortical atrophy associated with AD compared to that of NC. Due to its non-invasive property and relatively low cost, MRI is the commonly used neuroimaging modality in the diagnosis of AD.

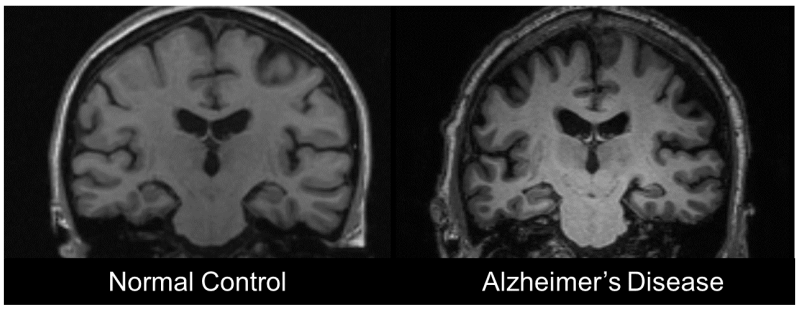


Figure 1.4: Differences between NC and AD captured by MRI scan (coronal view), from Alzheimer's Disease Neuroimaging Initiative (ADNI)

1.2.2 Functional magnetic resonance imaging

After the brain is stimulated, neuronal activation, regional cerebral blood flow and oxygen consumption will change. Due to this fact, functional MRI (fMRI) can measure brain activity via detecting changes associated with blood oxygen level dependent (BOLD) signal. If a brain region is used, blood flow to that region will increase [16]. fMRI mainly includes two categories: resting state fMRI (rs-fMRI) and

task state fMRI (ts-fMRI). Rs-fMRI is considered to be a potential modality for AD as functional brain changes are thought to precede structural brain changes [17]. Figure 1.5 [18] illustrates the lack of connectivity caused by AD. In the subject under NC, resting activity in the posterior cingulate seed region (marked by a star) is correlated with activity in inferior parietal and medial frontal regions, while in the AD subject, the long-range functional connectivity of the posterior cingulate seed region is greatly reduced, particularly with respect to the medial frontal cortex.

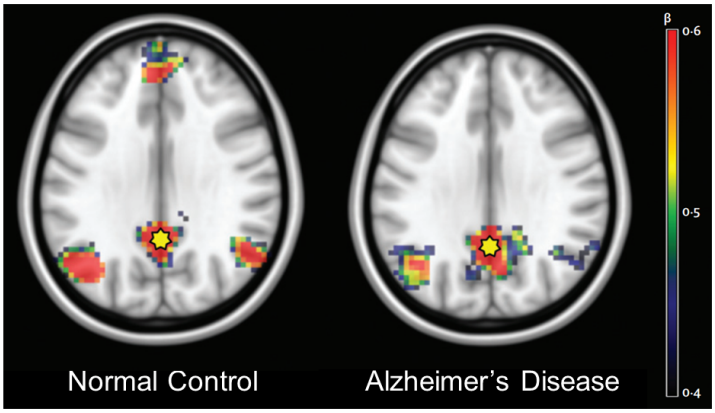


Figure 1.5: Differences of functional connectivity between NC and AD captured by rs-fMRI (axial view).

1.2.3 Diffusion tensor imaging

Diffusion tensor imaging (DTI) uses the dispersion anisotropy of water molecules for imaging and can be used for white matter fiber research. The structural basis of the cortical connection is the white matter fiber tracts between the cortex, and AD is currently considered to be a progressive cortical disconnection syndrome. DTI studies [19–21] have found that MCI patients have many white matter regional damages, such as frontal lobe, temporal lobe, parietal lobe and superior longitudinal fasciculus. Similar to functional connections, changes in structural connectivity during AD progression are earlier than significant gray matter shrinkage. These results suggest that DTI may be a modality for early diagnosis of AD. Fractional anisotropy (FA) is one of the useful indicators derived from the diffusion tensor that is closely related to white matter integrity [22]. As shown in Figure 1.6, the abnormal connections of white matter in temporal lobe reveal differences between the patient with AD and the subject under NC.

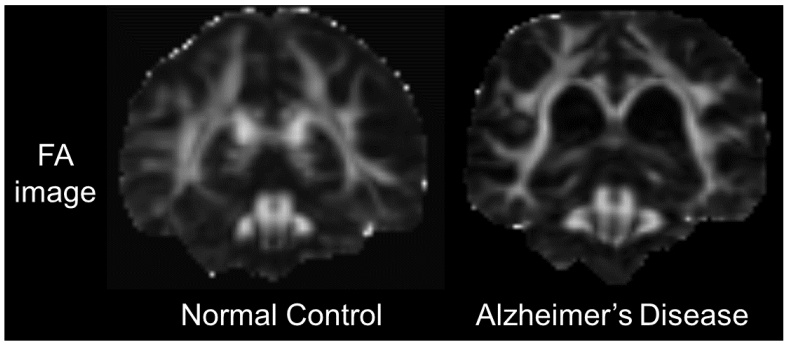


Figure 1.6: Differences in FA (fractional anisotropy) image between NC and AD captured by DTI (coronal view), from ADNI.

1.2.4 Single photon emission computed tomography

Single photon emission computed tomography (SPECT) is an in vivo neuroimaging technique that uses gamma-emitting radiotracers to assess cerebral blood flow [23]. The mainly used tracers include technetium-99m-labeled compounds, such as ^{99m}Tc-hexamethyl propylene amine oxime (^{99m}Tc-HMPAO) and ^{99m}Tc-ethyl cysteinate dimer (^{99m}Tc-ECD) [24]. These tracers are lipophilic and freely cross the blood-brain barrier in a manner proportional to the cerebral blood flow [23]. A single photon is emitted in a blood-rich brain tissue, and then using tomography and image reconstruction to form multiple azimuth sections and three-dimensional images. The changes in brain function are reflected by the measure of regional cerebral blood flow (rCBF). The majority of SPECT studies [25–27] have validated that evident deficits in perfusion can be observed in temporal and parietal regions in AD as compared to NC when using ^{99m}Tc-HMPAO and ^{99m}Tc-ECD as radiotracers, as illustrated in Figure 1.7 where the decreased perfusion in the parietal lobe can be clearly seen. In addition, the reduction of rCBF in the posterior cingulate gyrus of AD patients conduces to the early diagnosis of AD, and can also be used to predict the conversion from MCI to AD [28].

1.2.5 Positron emission tomography

Positron emission tomography (PET) is a nuclear medicine functional imaging method that uses a radiotracer to detect activities in the body and obtain cellular activities or metabolic information so as to aid diagnose of diseases [29]. There are different types of radiotracers that can be used in PET scanning, such as fluorodeoxyglucose (^{18}F-FDG), ^{11}C-Pittsburgh compound B (^{11}C-PiB) and ^{1}8F-

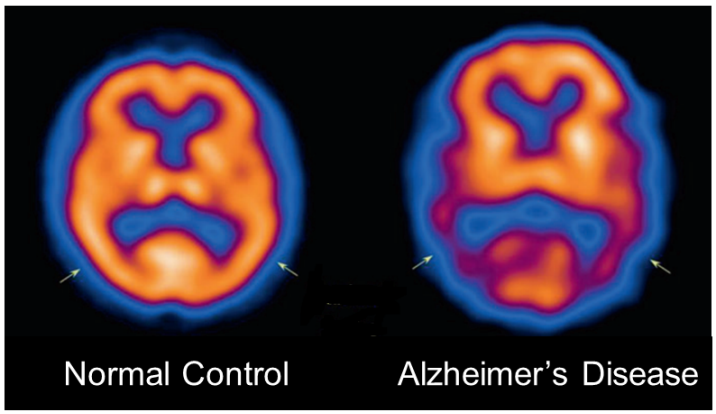


Figure 1.7: Differences between NC and AD captured by SPECT (axial view), from ADNI.

florbetapir (^{18}F-AV-45), *etc.*

The basic procedure for a PET scan involves injecting the patient with a radiotracer, and then scanning it after a short duration of injection. The duration depends on the type of radiotacer, and typically it lasts 30 min, 50 min and 50 min for ^{18}F-FDG, ^{11}C-PiB and ^{18}F-AV-45, respectively. During the positron emission decay, the radiotracer emits a positron which travels only a short distance through tissue and meanwhile loses kinetic energy until it is almost at rest. When this low energy positron interacts with an atomic electron, the particles annihilate to produce two gamma-ray photons that are detectable outside the body. To conserve energy and momentum, the photons must be emitted in opposite directions. Since the elements of the PET detector form closed rings around the patient, the two photons are detected simultaneously in opposite detector elements. This process, known as coincidence detection, allows spatial localization of the tracer in the body and the production of an image showing its distribution. The tissue or lesion with high metabolic rate has a clearly high or bright signal on PET, and vice versa.

Amyloid beta (Aβ) deposit is one of the hypotheses that causes AD. Amyloid PET, which uses radiotracer to detect the distribution of amyloid plaques in the brain, has a high agreement rate with autopsy results and can be used as a direct diagnostic marker for pathological changes in Aβ [30]. The currently used radiotracers for Amyloid PET imaging are mainly ^{11}C-PIB, ^{18}F-AV-45, ^{18}F-florbetaben and ^{18}F-flutemetamol. These radiotracers are useful to detect the cortical amyloidosis [31]. Figure 1.8 and Figure 1.9 illustrate differences of Aβ depositions between an NC subject and an AD patient in PiB-PET and Florbetapir-PET, respectively (without normalization). As can be seen, the AD patient has increased PiB or florbe-

tapir retention in regions known to accumulate significant Aβ deposits in comparison with the NC subject. Besides AD, Aβ deposits in the brain are also present in other neurodegenerative diseases associated to dementia, such as Parkinson's disease and dementia with Lewy bodies [31].

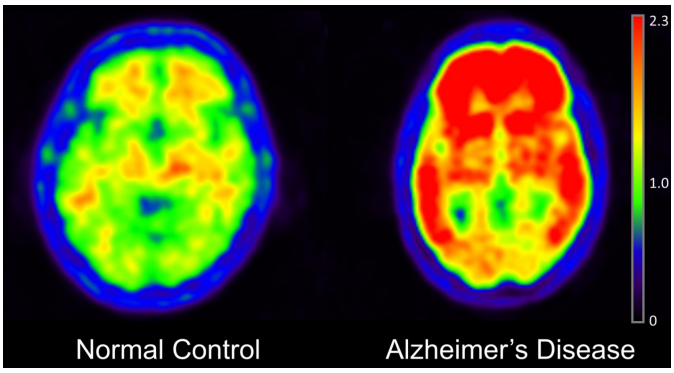


Figure 1.8: Differences between NC and AD captured by PiB-PET (axial view), from ADNI.

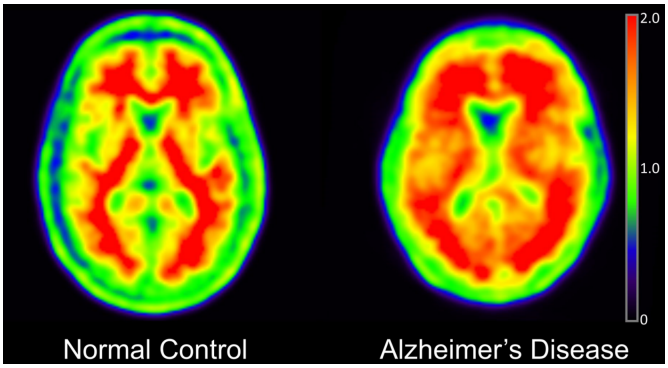


Figure 1.9: Differences between NC and AD captured by Florbetapir-PET (axial view), from ADNI.

^{18}F-FDG PET (referred to FDG-PET hereafter) is used to measure glucose uptake in neurons and glial cells and is considered to be a sensitive indicator of changes in synaptic function. Patients with AD had severe glucose metabolism defects, therefore the specific parts related to AD can be detected or located through the radiotracer, ^{18}F-FDG, which reflects the glucose metabolism. A number of studies on FDG-PET [32–35] have shown that patients with AD show locally significant low-metabolism on the overall low-metabolism background of the brain, including temporal lobe, parietal lobe, posterior cingulate gyrus, and expansion to frontal lobe as the disease progresses, compared with the normal age group, while the central anterior and posterior gyrus, cerebellum and thalamus are relatively normal. Moreover, changes in metabolic activity are believed prior to structure atrophy, therefore

FDG-PET is usually considered to be one of the effective modalities for early diagnosis. Figure 1.10 shows an instance in which AD patient can be distinguished from NC by using FDG-PET scan. It can be seen that the AD patient has typically reduced glucose metabolism in parietal lobe.

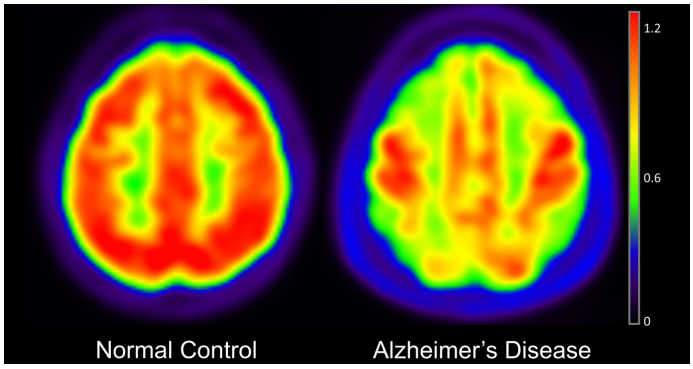


Figure 1.10: Differences between NC and AD captured by FDG-PET (axial view), from ADNI.

Table 1.1 shows different types of neuroimaging modalities relevant to AD. In clinical, it is difficult to diagnose the early stage of AD by a single test. Multiple tests are suggested to apply in order to provide information from multiple views.

Table 1.1: Neuroimaging modalities for AD diagnosis.

Modality	Discriminative Pattern	Abnormality
MRI	Regional volume	Atrophy/Reduced
fMRI	Functional connectivity	Reduced
DTI	Structural connectivity	Reduced
SPECT	Perfusion profile	Reduced
Amyloid PET	Amyloid plaques	Deposit/Increased
FDG-PET	Metabolism	Reduced

1.3 Contributions and outline

Neuroimaging test is necessary for AD diagnosis since neuroimaging modalities can capture changes in a brain. But if changes are subtle or slight, it becomes difficult to identify. Computer-aided diagnosis (CAD) methods can be designed to interpret medical images accurately. CAD is a set of methods that process images of typical appearance and highlight significant portions, such as possible diseases,

to provide support for decision making, thereby assisting doctors to analyze and evaluate comprehensively in a short period of time. Considering that metabolism changes is a key factor for early diagnosis of AD and MCI convention prediction, therefore, in this thesis, we focus on using CAD methods to investigate the distinctive patterns in FDG-PET data which can contribute to AD diagnosis and MCI conversion prediction.

In Chapter 2, two parts involved in CAD methods are presented respectively, including feature extraction and classification. Feature extraction is firstly introduced, which mainly covers the commonly used and novel features, and the corresponding feature reduction or selection techniques. Then an overview of machine learning concepts relevant to classification is provided, including classifiers, such as Logistic Regression (LR), Support Vector Machine (SVM), and evaluation metrics with which to assess the classification performance, such accuracy, sensitivity, specificity and area under curve. In addition, deep learning models are described, including MultiLayer Perceptron (MLP) and Convolutional Neural Network (CNN), followed by applications to AD diagnosis.

Most neuroimage-based AD classification methods are usually tested on a public dataset, Alzheimer's Disease Neuroimaging Initiative (ADNI) data. However, due to the complexity and diversity of the data, these studies have only exploited a subset and few of them have clarified the details of data selection. In Chapter 3, we clarify the procedure of data acquisition and the rule of data selection in order to provide guidance for other methods.

In Chapter 4, multilevel feature representation for FDG-PET data is investigated to diagnose AD and its early stage. The major contributions gained in this chapter can be summarized as three folds: 1) the multi-level feature representation considers not only region properties, but also the connectivity between any pair of regions and an overall connectivity between one region and the other regions; 2) a similarity-driven ranking method is proposed to rank regions from highly affected to slightly affected by the disease, which can reduce the feature dimension and increase the classifier's diversity to a certain degree; 3) a classifier selection strategy is proposed to choose a pair of classifiers with high diversity to enhance the ensemble effect, especially for the case that sub-classifiers do not perform well.

In Chapter 5, the features used to characterize FDG-PET images are extracted from another point of view—spatial gradients of FDG rates in PET brain images, instead of voxel-wise and ROI-wise features as many studies have done previously. This work is motivated by the observed differences of FDG rates between AD and

NC subjects. The spatial gradients are quantified by a 2D histogram of orientation, which is similar to Histogram of Oriented Gradient (HOG) [36] that has been successfully applied for object detection in 2D images. The contributions involved in this chapter can be summarized into three aspects: 1) 1D HOG descriptor, used in natural scene images originally, is improved to 2D HOG to quantify spatial gradients, thereby characterizing 3D FDG-PET brain images. Moreover, 2D HOG is expressed in a multiple scale manner, which proves to be more effective than the commonly used features; 2) a region ranking method is proposed to select distinctive ROIs by using multiscale HOG features; 3) an ensemble classification framework is designed through considering different scales of HOG descriptors for the individual region and concatenated regions, which enhances the diagnosis accuracy for the classification.

As the impressive performance has been gained by deep learning, in Chapter 6, we attempt to diagnose AD and predict MCI conversion under the framework of CNN. Accordingly, two multiview CNN architectures are proposed, denoted mvCNNiF and mvCNNaF, with differences in the combination manner of multiple views, including axial, coronal and sagittal views. Apart from this, another main contribution of this chapter lies in a mapping layer with cuboid convolutional kernels which is designed to project information along the third dimension onto a plane. The proposed mapping layer can reduce parameters involved in the following convolutional layers and meanwhile consider spatial relations in different views.

Chapter 7 summarizes the thesis, in particular for the three developed methods, and then proposes the future work from three aspects, including data, methods and tasks.

Chapter 2

Computer-aided Diagnosis Methods: Application to Alzheimer's Disease

2.1 Introduction

Computer-aided Diagnosis (CAD) methods generally consist of two parts, one part is feature extraction and the other one is related to classification, as illustrated in Figure 2.1. In the following sections, we mainly describe CAD methods applied to Alzheimer's disease (AD) diagnosis and its corresponding early stage, Mild Cognitive Impairment (MCI), under the modality of FDG-PET. In Section 2.2, discriminative features for characterizing FDG-PET images are presented, as well as methods of feature reduction or selection. Then details of various classification algorithms are described, Section 2.3 is for classical machine learning methods and Section 2.6 is for deep learning techniques.

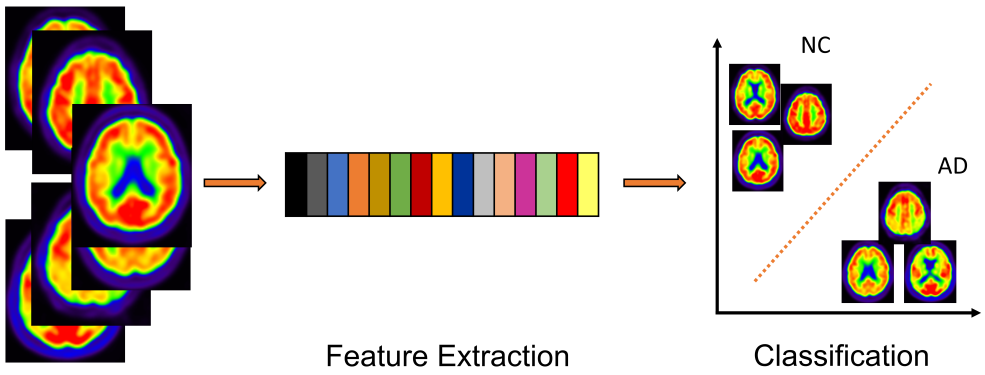


Figure 2.1: CAD framework for Alzheimer's disease.

2.2 Discriminative features for representations

Feature extraction is a process of using domain knowledge to create features in order to make machine learning algorithms work. Such features are called hand-crafted features compared to the features automatically learned by deep learning models. In this section, we mainly describe hand-crafted features that can represent FDG-PET data.

2.2.1 Voxel-wise features

FDG-PET images reflect the glucose uptake rate, and lower intensities indicate lower metabolic activities, such activities are usually considered abnormal. Therefore, voxel intensities can be viewed as a kind of feature to characterize FDG-PET images, as has been done in [37–39], these studies used voxel intensities of FDG-PET images as features to classify AD from NC. However, medical images are typically high-dimensional, although all the voxels are used to retain all the information, feature redundancy is difficult to avoid, which makes the training procedure time-consuming. For example, an FDG-PET image with a size of $91 \times 109 \times 91$, its voxel number in gray matter is $\sim 10^6$. In order to address this problem, ROI-wise features could be used.

2.2.2 ROI-wise features

ROI-wise feature extraction is under the help of a pre-defined atlas, through which a whole brain can be segmented into different Regions of Interest (ROI). Then statistical features of each ROI are computed, such as region's mean intensity and standard deviation. Consequently, the feature dimension is reduced to $\sim 10^2$ or $\sim 10^3$, which depends on the atlas. There is variety of atlas used in brain segmentation, including Anatomical Automatic Labeling (AAL) [40] which could be the most widely used parcellation approach, and its updated version AAL2 [41] which refined the parcellation of orbitofrontal cortex on the basis of AAL, LONI Probabilistic Brain Atlas (LPBA40) [42], Hammers_mith [43, 44], and other atlases. Different methods yield different numbers of ROIs, for example, AAL atlas is composed of 116 regions (90 cerebral regions and 26 cerebellar regions) and AAL2 atlas consists of 120 regions (4 more cerebral regions), whereas LPBA40 has 56 regions (54 cerebral regions, brainstem and the cerebellum), Hammers_mith atlas has a variety of releases, including 67 regions [45], 83 regions and 95 regions [46]. In addition, a

specific atlas can be created according to data by using clustering methods. Li *et al.* [47] created their own atlas through a Gaussian Mixture Model (GMM) to group brain voxels into small regions. For now, there is not an atlas that is absolutely effective. It is a complex task to choose an appropriate atlas since it is associated with multiple factors, such as raw data, data type, data processing and the task it will be applied to. Samper-Gonzalez *et al.* [48] has reported that no atlas consistently outperformed others across tasks. For instance, in [48], LPBA40 atlas performed better than AAL2 and Hammers_mith atlases under MRI modality for distinguishing AD from NC, while AAL2 atlas was more effective by using FDG-PET modality for the same task. Therefore, we need to trade off performance on different modalities and tasks in order to select a more suitable atlas. AAL atlas is applied in this thesis since it is widely used in FDG-PET analysis and yields good results [49–54].

Generally, ROI-wise features are most commonly used in AD identification involving FDG-PET modality because they do reflect distinct characters between normal and abnormal subjects and meanwhile have a lower feature dimension compared to voxel-wise features. Most studies [49, 55–57] have exploited mean intensity of each region as the ROI-wise feature. For example, Panani *et al.* [49] applied AAL atlas to segment each FDG-PET image into 90 ROIs and used regional features to predict MCI conversion. Gray *et al.* [55] used regional intensities to analyze longitudinal FDG-PET data in AD classification. Both Ota *et al.* [56] and Asim *et al.* [57] have used regional average intensity as the feature to compare performance of AAL atlas and LPBA40 atlas in AD discrimination under FDG-PET modality. In addition to mean intensity, other regional parameters, such as standard deviation, entropy, are applied to address the problem to AD diagnosis as well. Li *et al.* [47] and Garali *et al.* [52] have explored the effects of regional standard deviation in AD recognition. Also, Garali *et al* [53] have used regional skewness, kurtosis and entropy, combined with regional mean intensity and variance to classify AD from NC, which has achieved better results than regional mean value features in their local dataset. Furthermore, regional text features, such as Gray-level co-occurrence matrix (GLCM), Gray-level size zone matrix (GLSZM), and wavelet features can also be applied and have obtained competitive results for three tasks, including AD vs. NC, MCI vs. NC and AD vs. MCI [54].

2.2.3 Feature selection/reduction

As mentioned above, ROI-wise feature could be viewed as a way to reduce the dimension of voxel-wise feature. In fact, not only voxel-wise features but also ROI-wise features may have the problem of feature redundancy. To address the problem, feature selection as well as feature reduction can be applied. Generally, feature selection is referred to select a subset of raw features, while feature reduction means to transform raw features to a lower-dimensional space. Feature selection methods could be categorized into three groups: filter, wrapper and embedded methods [58].

Filter

Filter methods use a pre-defined measurement to rank features and then select top K features or select features through setting a threshold. The measurements could be variance, Pearson Correlation Coefficient (PCC) [59], mutual information [60] and scores of significance tests, such as t-test, χ^2-test. These criteria measure the dispersion of samples under different features, such as variance, or the correlation and dependence between observations and the corresponding respondents, such as Pearson correlation coefficient and mutual information. In addition to metrics, using prior knowledge to select features can also be categorized to filter methods. The basic idea behind these methods is to select features, especially anatomical regions through pre-acquired knowledge. Teipel *et al* [61] have selected 42 regions among 83 regions obtained through Hammers_mith atlas because the removed regions are known to be not prominently involved in AD, such as the cerebellum and the ventricles.

Wrapper

Wrapper methods use classification models to test the performance of a given subset of features and the performance is usually evaluated through Area Under Curve (AUC), Mean Squared Error (MSE) or accuracy. The feature set with a high metric value will be selected. Since the wrapper methods train a new model for each subset, they are computationally intensive but typically provide the best performing feature set for a particular type of model.

Embedding

Embedding methods integrate feature selection into the model construction and they tend to be between filters and wrappers in respect of computational complexity.

A typical example of this method is least absolute shrinkage and selection operator (LASSO) algorithm [62]. Give n samples, $\mathbf{x}_1$, $\mathbf{x}_2$,...,$\mathbf{x}_i$,..., $\mathbf{x}_n$, and $\mathbf{x}_i$ is a vector which contains t values. Each element in $\mathbf{x}_i$ can be considered as a feature. The corresponding outputs are y_1, y_2,...,y_i,..., y_n, respectively. The objective of lasso is to solve the problem,

$$\begin{aligned}(\beta_0{}^*, \boldsymbol{\beta}^*) =& \arg\min_{\beta_0,\boldsymbol{\beta}} \left(\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 \right) \\ & \text{s.t.} \sum_{j=1}^{t} |\beta_j| \leqslant c \end{aligned} \tag{2.1}$$

where parameters β_0 and $\boldsymbol{\beta}$ are a scalar and a regression coefficient vector of length t, respectively, c is a pre-specified free parameter that determines the amount of regularization. Any feature with a non-zero regression coefficient will be selected by LASSO algorithm, in other words, if β_j equals zero, then the j-th feature in $\mathbf{x}_i$ will be eliminated. There are other improved versions on the basis of lasso, such as ridge regression [63] which replaces ℓ_1 regularization with ℓ_2 regularization, Elastic Net [64] which combines the two regularization methods. It should be noted that ridge regression does not have the effect of selecting features.

Feature reduction is a kind of projection method which projects the higher-dimensional feature space into a lower-dimensional space, thereby achieving the goal of dimensionality reduction. The commonly used techniques are Principal Components Analysis (PCA) [65] and Linear Discriminant Analysis (LDA) [66]. For a given set of data with n dimensions, both PCA and LDA aim to find a subspace of dimension d lower than n such that the data can be mapped onto this subspace, but PCA chooses the projection direction in which the projected data have the largest variance, while LDA attempts to choose the direction that enables the two classes to have the largest difference [67]. Besides, Non-negative Matrix Factorization (NMF) can be classified into this category if used for feature reduction. Padilla *et al.* [68] applied NMF to select discriminative voxels in AD diagnosis.

Table 2.1 lists the studies using different feature types and feature selection/reduction methods. It can be seen that most approaches utilized the feature selection or reduction strategy. An appropriate strategy can not only reduce training complexity, but also improve the performance.

Table 2.1: A brief overview of features used in CAD methods for Alzheimer's disease.

Method	Modality	Feature type	Atlas(#ROIs)	Feature selection/reduction
Vandenberghe *et al.* [37]	FDG-PET	Voxel-wise	None	None
Gray *et al.* [38]	MRI & FDG-PET	Voxel-wise (for FDG-PET)	None	None
Tong *et al.* [39]	MRI & FDG-PET	Voxel-wise (for FDG-PET)	None	None
Salas *et al.* [69]	FDG-PET	Voxel-wise	None	t-test
Padilla *et al.* [68]	FDG-PET	Voxel-wise	None	NMF
Hinrichs *et al.* [70]	FDG-PET	Voxel-wise	None	t-test
Cabral *et al.* [71]	FDG-PET	Voxel-wise	None	Mutual information
Illán *et al.* [72]	FDG-PET	Voxel-wise	None	PCA
Pagani *et al.* [49]	FDG-PET	ROI-wise	AAL (90)	Prior knowledge
Li *et al.* [47]	FDG-PET	ROI-wise	clustering	Accuracy-driven
Garali *et al.* [52]	FDG-PET	ROI-wise	AAL(116)	AUC-driven
Garali *et al.* [53]	FDG-PET	ROI-wise	AAL (116)	AUC-driven
Li *et al.* [54]	FDG-PET	ROI-wise	AAL (116)	t-test & PCC
Gray *et al.* [55]	FDG-PET	ROI-wise	Hammers_mith (83)	None
Teipel *et al.* [61]	MRI & AV45-PET & FDG-PET	ROI-wise	Hammers_mith (42)	Prior knowledge
Zhang *et al.* [73]	MRI & FDG-PET	ROI-wise	other [74] (93)	t-test
Shi *et al.* [75]	MRI & FDG-PET	ROI-wise & inter-ROI	other (93)	Lasso
Zu *et al.* [76]	MRI & FDG-PET	ROI-wise	other [74] (93)	Embedding
Zhu *et al.* [77]	MRI & FDG-PET	ROI-wise	other [74] (93)	Customized

2.3 Machine learning techniques for classifications

Machine learning is a branch of artificial intelligence involving segmentation, regression, detection, clustering, classification, *etc.* The basic idea of "learning" is to automatically improve the performance with experience which can be seen as patterns and inference. The focus of the thesis is medical diagnosis which is a problem associated to classification. There is variety of classifiers, such as logistic regression [78], decision trees (CART [79], ID3 [80], C4.5 [81]), neural network [82], Support Vector Machine (SVM) [83], naive Bayes classifier [84] and ensemble methods (boosting and bagging). As logistic regression is the basis of neural network and SVM combined with ensemble methods is exploited in the thesis, therefore in the following parts, we mainly describe the three kinds of methods. Before introducing different classifiers, we firstly describe the problem to be solved by classification from the point of mathematics view.

Still considering n samples $\mathbf{x}_1$, $\mathbf{x}_2$,...,$\mathbf{x}_i$,..., $\mathbf{x}_n$, like mentioned in Section 2.2.3, and each sample can be characterized by t-dimensional feature vector, $\mathbf{x}_i = (x_{i1}; x_{i2}; \ldots; x_{it})$. The corresponding respondent of $\mathbf{x}_i$ is y_i, thus an arbitrary dataset can be denoted $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)$. The objective of classification is to construct a mapping from input space $\mathcal{X}$ to output space $\mathcal{Y}$, f: $\mathcal{X} \mapsto \mathcal{Y}$, through learning from the dataset. For the binary classification, $\mathcal{Y} = \{-1, +1\}$ or $\{0, 1\}$. For the multi-classification, $|\mathcal{Y}| > 2$ ($|\cdot|$ denotes the cardinality of a set). For regression, $\mathcal{Y} = \mathbb{R}$, $\mathbb{R}$ is real numbers. A prediction can be made for a new sample after constructing the mapping, $y_{n+1} = f(\mathbf{x}_{n+1})$. The mapping function can be seen as the classification model.

2.3.1 Logistic Regression

A linear model attempts to learn a mapping function according to combining features linearly,

$$f(\mathbf{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_t x_t + b \tag{2.2}$$

and generally it is rewritten to its vector form,

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \mathbf{x} + b \tag{2.3}$$

where $\mathbf{w} = (w_1; w_2; \ldots; w_t)$ is a set of weights and b is a bias. Learning the mapping function consists in finding the weights and bias such that for $\mathbf{x}_i$, its output close to

the corresponding true value,

$$f\left(\mathbf{x}_i\right) \simeq y_i \tag{2.4}$$

and the least square method [85] is applied to estimate parameters $\mathbf{w}$ and b. Then we can get an objective function,

$$\begin{aligned}(\mathbf{w}^*, b^*) &= \arg\min_{(\mathbf{w},b)} \sum_{i=1}^{n} \left(y_i - f\left(\mathbf{x}_i\right)\right)^2 \\ &= \arg\min_{(\mathbf{w},b)} \sum_{i=1}^{n} \left(y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i - b_i\right)^2\end{aligned} \tag{2.5}$$

and its corresponding vector form is,

$$(\mathbf{w}^*, b^*) = \arg\min_{(\mathbf{w},b)} (\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{w} - b)^{\mathrm{T}}(\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{w} - b) \tag{2.6}$$

where $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n})$. Let $E(\mathbf{w}, \mathbf{b}) = (\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{w} - \mathbf{b})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{w} - \mathbf{b})$, and let derivatives of function E with respect to $\mathbf{w}$ and b equal to zero, respectively. Then we can get estimations of the two parameters. Since $\mathbf{w}$ visually expresses the importance of each feature in the prediction results, thus linear models have a good comprehensibility.

In fact, the prediction results of Eq. 2.3 are real numbers, and $f(\mathbf{x})$ is a linear regression model. For the binary classification, we need to find a monotonic differentiable function that can link true labels with regression results. Logistic function, one of sigmoid functions, can play that role,

$$y = \frac{1}{1 + e^{-z}} \tag{2.7}$$

as shown in Figure 2.2. Its significance is intuitive, that is, when the prediction z is

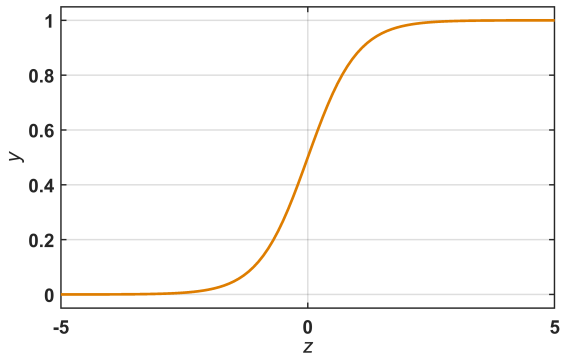


Figure 2.2: Logistic function.

positive, y will tend to 1, while z is negative, y tends to 0. z is $f(\mathbf{x})$ actually, thus substitute Eq. 2.3 into Eq. 2.7,

$$y = \frac{1}{1 + e^{-(\mathbf{w}^{\mathrm{T}}\mathbf{x}+b)}} \tag{2.8}$$

which is called logistic regression [78]. It solves classification problems despite being called regression. In order to simplify the computation, the bias term, b, is included into $\mathbf{w}$, which will lead to $\mathbf{x}_i \leftarrow (\mathbf{x}_i; 1)$ and $\mathbf{w} \leftarrow (\mathbf{w}; b)$. Consequently, Eq. 2.8 is rewritten to,

$$y = \frac{1}{1 + e^{-(\mathbf{w}^{\mathrm{T}}\mathbf{x})}} \tag{2.9}$$

Eq. 2.9 can be seen as the posterior probability estimation of predicting y as 1 when $\mathbf{x}$ is positive,

$$P(y = 1|\mathbf{x}; \mathbf{w}) = \frac{1}{1 + e^{-(\mathbf{w}^{\mathrm{T}}\mathbf{x})}} = p \tag{2.10}$$

thus the probability of predicting as 0 is,

$$P(y = 0|\mathbf{x}; \mathbf{w}) = 1 - p \tag{2.11}$$

Eq. 2.10 and Eq. 2.11 are equivalent to

$$P(y_i|\mathbf{x}_i; \mathbf{w}) = p^{y_i}(1 - p)^{1-y_i} \tag{2.12}$$

Suppose all the samples are independent and identically distributed, according to the Maximum Likelihood Estimation [86], the likelihood function is,

$$\begin{aligned} L(\mathbf{w}|\mathbf{x}, \mathbf{y}) &= P\left(y_1|\mathbf{x}_1; \mathbf{w}\right) P\left(y_2|\mathbf{x}_2; \mathbf{w}\right) \ldots P\left(y_n|\mathbf{x}_n; \mathbf{w}\right) \\ &= \prod_{i=1}^{n} p^{y_i}(1 - p)^{1-y_i} \end{aligned} \tag{2.13}$$

after taking logarithm for both sides, we can obtain

$$\begin{aligned} l(\mathbf{w}) &= \ln\left(\prod_{i=1}^{n} p^{y_i}(1 - p)^{1-y_i}\right) \\ &= \sum_{i=1}^{n} \left(y_i \ln p + (1 - y_i) \ln(1 - p)\right) \end{aligned} \tag{2.14}$$

which is the objective function of logistic regression model. The probability that each sample belongs to its true label is expected to be largest, thus the parameter $\mathbf{w}$ can be obtained through using the gradient descent method to optimize the following

equation.

$$\begin{aligned} \mathbf{w}^* &= \arg\max_{\mathbf{w}} l(\mathbf{w}) \\ &= -\arg\min_{\mathbf{w}} l(\mathbf{w}) \end{aligned} \tag{2.15}$$

2.3.2 Support Vector Machine

Support vector machine (SVM) classifier [83] aims to construct a hyperplane that maximizes the margin which is a distance between the closest points on either side of the boundary. These points are known as the support vectors. In the sample space, the hyperplane can be expressed as

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0 \tag{2.16}$$

where $\mathbf{w}$ is seen as the normal vector determining the direction, while the bias term b controls the distance between the origin and the hyperplane. Clearly, once $\mathbf{w}$ and b are confirmed, the hyperplane can be achieved.

The distance from an arbitrary sample to the hyperplane is,

$$D = \frac{\left|\mathbf{w}^{\mathrm{T}}\mathbf{x} + b\right|}{\|\mathbf{w}\|}. \tag{2.17}$$

Suppose a hyperplane can classify samples correctly, then let

$$\begin{cases} \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b \geqslant +1, & y_i = +1 \\ \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b \leqslant -1, & y_i = -1 \end{cases} \tag{2.18}$$

as shown in Figure 2.3 intuitively, the support vectors marked with filled color determine the margin,

$$D = \frac{2}{\|\mathbf{w}\|}. \tag{2.19}$$

The objective function is the maximization of the margin and it can therefore be expressed as a constrained optimization,

$$\begin{aligned} &\max_{(\mathbf{w},b)} \frac{2}{\|\mathbf{w}\|} \\ &\text{s.t. } y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\right) \geqslant 1, \forall i \in \{1, 2, ..., n\} \end{aligned} \tag{2.20}$$

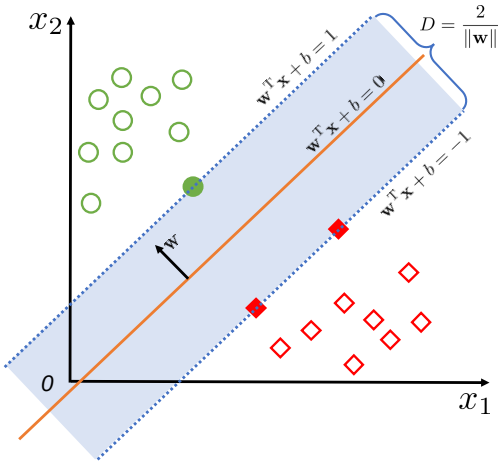


Figure 2.3: Maximum-margin hyperplane and margins for an SVM.

which is equivalents to

$$\begin{aligned} &\min_{(\mathbf{w},b)} \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{s.t. } y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\right) \geqslant 1, \forall i \in \{1, 2, ..., n\} \end{aligned} \tag{2.21}$$

where the constraint ensures that no feature vectors fall within the margin. Lagrange multipliers are applied to transform a constrained optimization problem to an unconstrained one,

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\right)\right) \tag{2.22}$$

where $\boldsymbol{\alpha} = (\alpha_1; \alpha_2; \ldots; \alpha_n)$ and α_i is a Lagrange multiplier. Let the differential of L to $\mathbf{w}$ and b be 0, we can have,

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \tag{2.23}$$

$$0 = \sum_{i=1}^{n} \alpha_i y_i \tag{2.24}$$

then substitute Eq. 2.23 and Eq. 2.24 into Eq. 2.22, a dual expression of Eq. 2.22 is computed,

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ &\text{s. t. } \alpha_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{2.25}$$

in which $\boldsymbol{\alpha}$ can be solved by Sequential Minimal Optimization (SMO) method [87], and then $\mathbf{w}$ and b are obtained.

In practical applications, data is not absolutely linearly separable and certainly contains misclassified instances. The problem can be addressed by the soft-margin SVM in which slack variables ξ are introduced to enable the classifier to deal with data that could not be completely separated, such as noise data. Therefore, the optimization becomes a trade-off between maximizing the margin and minimizing the degree of misclassification. This trade-off is controlled by the penalty parameter C, such that the constrained optimization is expressed as,

$$\begin{aligned} &\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{s. t. } y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\right) \geqslant 1-\xi_i \quad \text{and} \quad \xi_i \geqslant 0 \end{aligned} \tag{2.26}$$

Similarly, its dual formalization is,

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j \\ &\text{s. t. } 0 \leqslant \alpha_i \leqslant C \quad \text{and} \quad \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned} \tag{2.27}$$

where $\boldsymbol{\alpha}$ is limited under an upper bound C compared to Eq. 2.29.

A limitation of linear classifiers is that, when data is intrinsically nonlinear, they cannot separate them well. In such cases, a general approach is to map the data points onto a higher-dimensional feature space where the data linearly non-separable in the original feature space become linearly separable. For this purpose, SVM is improved by applying the kernel trick to maximum-margin hyperplanes [88] and the corresponding objective function is,

$$\begin{aligned} &\min_{(\mathbf{w},b)} \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{s.t. } y_i\left(\mathbf{w}^{\mathrm{T}}\phi\left(\mathbf{x}_i\right) + b\right) \geqslant 1 \end{aligned} \tag{2.28}$$

where $\phi\left(\mathbf{x}_i\right)$ is a mapped vector. The dual expression is,

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) \\ &\text{s. t. } \alpha_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned} \tag{2.29}$$

where $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^{\mathrm{T}} \phi(\mathbf{x}_j)$ and it is a kernel function which could be polynomial, Gaussian radial basis function (RBF) and Hyperbolic tangent. Among them, Gaussian radial basis function is mostly used.

2.3.3 Ensemble methods

Ensemble methods train multiple learners to solve the same problem. In contrast to ordinary learning approaches which attempt to construct one learner from training data, ensemble methods try to construct a set of individual learners and integrate them, as illustrated in Figure 2.4. The individual learners could be the same type or different types, for instance, all the learners can be decision trees, or one learner is logistic regression, another is decision trees and the third one may be SVM. To get performance, it is generally believed that individual learners should be as accurate as possible, and as diverse as possible [89].

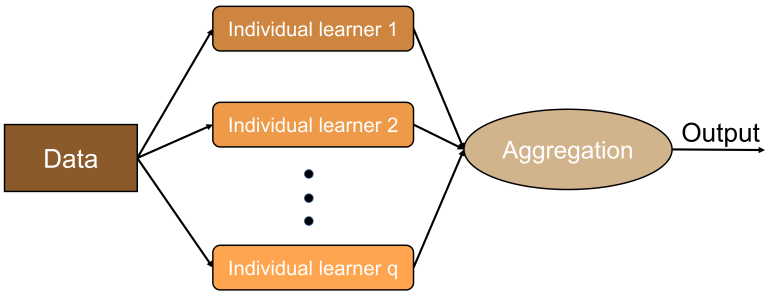


Figure 2.4: An ensemble architecture.

The accuracy depends on discriminative features and a classifier's intrinsic property. In the case that features have been extracted and the types of individual learners are fixed, the performance of an ensemble classifier can be improved by increasing the diversity among individual learners. The common basic idea for increasing diversities is to inject some randomness into the learning process. Popular mechanisms include manipulating the data samples, input features, and training parameters. It should be noticed that different mechanisms for diversity increment can be used at the same time.

Data Sample Manipulation is a commonly used mechanism. Given a dataset, multiple different subsets of data can be generated through sampling approaches, and then the individual learners are trained from different data subsets.

Input Feature Manipulation. The training data is usually characterized by a set of features. Different subsets of features can provide different views on the data. Therefore, individual learners trained from different subsets of features are usually

diverse. For data with a lot of redundant features, training a learner in a subset will be not only effective but also efficient.

Training Parameter Manipulation attempts to generate diverse individual learners through setting different parameters in the training stage. For example, different penalty parameters can be applied to individual SVM, or different initial weights can be assigned to individual neural networks.

The aggregation strategy is to weight the importance of different individual learners and then combine them according to the importance, which can also affect the performance of the ensemble classifier. Three main types of aggregations are usually applied, including averaging, voting and learning.

Averaging could contain two kinds of methods, simple averaging and weighted averaging. Simple averaging obtains the combined output by averaging the outputs of individual learners directly, while weighted averaging obtains the combined output by averaging the outputs of individual learners with different weights implying different importance. Obviously the simple averaging method is a special case of the weighted one, which means individual learners have the same significance. Actually, there are a lot of weights to be trained for a large ensemble classifier, which can easily lead to overfitting. Thus, weighted averaging is not absolutely better than simple averaging. It is generally accepted that simple averaging is suitable for integrating learners with similar performance, and if individual learners exhibit large differences, weighted averaging of unequal weights may achieve better performance.

Voting is a popular and fundamental aggregation method and also includes two categories, majority voting and plurality voting. Specifically, for majority voting, each individual classifier votes to select a label, and the final output is a label that gets more than half of the votes; if no class label gets more than half of the votes, a rejection option will be given and the ensemble classifier makes no prediction. In contrast to majority voting which requires at least half of votes, plurality voting takes the class label which receives the largest number of votes as the final prediction.

Learning is a method that another learner is trained to combine individual learners. Staking [90, 91] is a typical representative of this method. The basic idea is to train individual learners using the original training data set, and then their outputs and original labels are used to train a new learner.

According to the way of training individual learners, the ensemble methods can be roughly divided into two categories, 1) individual learners that have strong dependencies are trained in a serialized way, such as boosting [92]; 2) individual learners without strong dependency are trained in parallel, such as bagging [93].

Boosting is an ensemble method designed to generate a single strong classifier by combining multiple weak classifiers. A weak learner refers to a learner with slightly better performance than a random guess, while a strong classifier can usually achieve good performance. Specifically, an individual learner is firstly trained from the initial training set, and then the training samples' distribution is adjusted according to the performance of that learner, so that the training samples that are misclassified by the previous learner receive more attention in the following. The next individual learner is then trained based on the adjusted sample distribution. This is repeated until the number of individual learners reaches a predefined value, and these learners are finally integrated with using a weighted summation. AdaBoost [94] is a classic and popular boosting method.

Bagging is a parallel ensemble method which combines individual learners with high diversity. As mentioned above, the diversity among individual learners is crucial for enhancing the ensemble performance, and data sample manipulation is one of the methods to achieve that purpose. In contrast to generate multiple non-overlapped data subsets, bagging exploits bootstrap sampling [95] to generate different overlapped dataset for training individual learners. In detail, given a dataset containing n samples, a sample is randomly taken into the subset at first, and then put the sample back into the initial dataset so that the sample may still be selected at the next sampling. In this way, after n random sampling operations, we get a subset containing n samples in which some samples appear several times, and some never appear. By repeating the process q times, q data subsets of n samples are obtained. Therefore, we can train q individual learners in total. In order to aggregate these learners, Bagging adopts voting for classification and averaging for regression. Random forest (RF) [96] is a kind of improved Bagging. The main difference between Bagging is the incorporation of random feature selection.

2.4 Performance evaluation

In real tasks, there are usually a number of alternative learning models to choose among as well as several parameters to tune. We need to select the model and its corresponding parameter setting with the best performance. In order to achieve this goal, performance evaluation is necessary, which involves the commonly used and accepted measures and experiment designs.

2.4.1 Evaluation measures

For a binary classification task, we can obtain different combinations of true labels and predicted labels, that are, true positive (TP) representing correctly identified positive labels, true negative (TN) representing correctly identified negative labels, in contrast to those incorrectly classified, false positive (FP) which implies negative labels incorrectly classified as positive ones and false negative (FN) implying positive labels incorrectly classified as negatives. Those concepts can be expressed intuitively by a confusion matrix, as shown in Table 2.2 in which the summation of TP, TN, FP and FN is equal to the total number of samples.

Table 2.2: A confusion matrix for a binary classification.

True label	Predicted label	
	positive	negative
positive	TP	FN
negative	FP	TN

The mainly used measures for testing the performance of a binary classifier are computed based on the confusion matrix, including accuracy (ACC) or balanced accuracy (bACC) [97], sensitivity (SEN), specificity (SPE) and area under curve (AUC) which is usually referred to a Receiver Operating Characteristic (ROC) curve [98].

Accuracy is the proportion of samples that are correctly predicted, which is computed through,

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.30}$$

Sensitivity implies the proportion of correctly classified patients and it is expressed as,

$$\text{SEN} = \frac{TP}{TP + FN} \tag{2.31}$$

which is also known as true positive rate (TPR).

Specificity means the proportion of control samples that are correctly classified, which is defined as,

$$\text{SPE} = \frac{TN}{TN + FP} \tag{2.32}$$

and 1−SPE is referred to as false positive rate (FPR).

However, for the case that the dataset is unbalanced, it would yield an optimistic evaluation if ACC is used as a measure to test the classifier's performance. Because

the classifier would be biased and dominated by the majority data in that task. To address this problem, **balanced accuracy** is introduced and it is expressed as,

$$\text{bACC} = \frac{SEN + SPE}{2}. \tag{2.33}$$

ROC curve reflects the relationship between the true positive rate (SEN) and the false positive rate (1−SPE) with the change of the discrimination threshold of a binary classifier, as shown in Figure 2.5. An ideal classifier would achieve 100% sensitivity and specificity, as a result, the upper left corner is with the best performance. Generally the metric AUC is applied to quantify the ROC and a higher value indicates better performance.

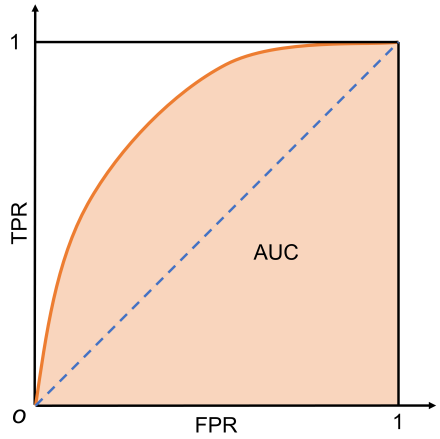


Figure 2.5: An instance of ROC curve.

2.4.2 Cross-validation

A classifier is optimized based on the training dataset. An independent test set is therefore required to assess that classifier in terms of its application and generalization to new unknown data. If there is not enough labeled data available to create such a test set separately, a common method is to use the cross-validation (CV) technique [99, 100] in which the test set is also known as the validation set. There are roughly three types of CV methods, hold-out, k-fold, leave-one-out (LOO). **Hold-out CV** is easy to understand, which randomly divides the dataset into two groups, one is used for training, the other one is as the validation set. Note that the class percentage should be maintained for both training and validation sets when splitting the data, which is called stratification or stratified sampling. Obviously, the hold-out method only involves one round validation. It is normal to use such

a method when the dataset contains a huge number of data since validation set could include a proper number of data to test the performance and meanwhile the computation consumption would be dramatically reduced. For instance, ImageNet [101], a very popular dataset for visual recognition or object detection, contains over 14 million images for now. However, for the case that the dataset is not that large, which is common in medical data, hold-out method is not a suitable way to evaluate the classifier's performance because it is highly possible to achieve misleading results due to the small validation set. To solve this problem, k-fold method is commonly used.

K-fold CV takes advantage of different partitions of the dataset and performs multiple rounds of validation, in such way, a more accurate estimation of model prediction performance will be achieved since it combines multiple results. In details, a dataset is usually partitioned with stratification into k equal-size subsets, then each subset could be the validation set and the others are the training set. Consequently, there are k combinations of training-validation set, and each of them is performed with the training-testing operation. After that, we can get k independent results, their average result is taken as the result of the cross-validation. Figure 2.6 gives an insight into the procedure of cross-validation. To reduce the influence of randomness caused by data splitting, the k-fold CV is usually repeated multiple times, and the common configurations, including 10-times 10-fold CV and 5-times 2-fold CV, are suggested in [102].

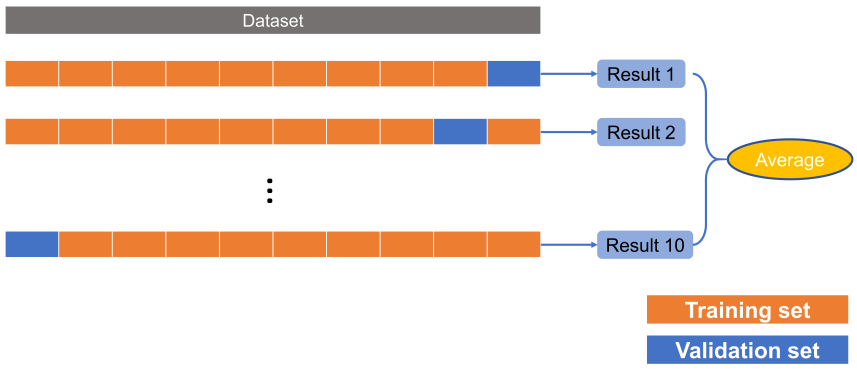


Figure 2.6: An instance of cross-validation.

LOO CV is a special case of k-fold CV in which k equals the number of samples in the original dataset and there is only one instance in each validation set. This method is usually applied when the dataset is particularly small, such as less than 50 samples.

2.5 Application to Alzheimer's disease diagnosis

Classifiers can tackle the problem of AD diagnosis automatically, among of them, SVM has been widely and successfully used, not only for FDG-PET data [53, 55, 68, 71], but also for other modalities, such as MRI [103–105], fMRI [106, 107], DTI [108, 109], or multi-modality [39, 61, 75, 76]. In recent studies on FDG-PET, reported accuracies are generally within the range of $[80\%, 96\%]$ for the task of AD diagnosis, while for pMCI prediction (progressive MCI vs. stable MCI), accuracies are within the range of $[60\%, 85\%]$. For the task of AD vs. NC, the highest accuracy, 95.95%, is achieved by Zu *et al.* in [76]. They applied a multi-kernel SVM to fuse the selected features from multi-modality data (MRI and FDG-PET) for final classification. They also tested their proposed method in the task of pMCI vs. sMCI and obtained an accuracy of 69.78%, which is inferior to Cabral's method [71] in which a higher accuracy, 85%, is obtained. Cabral *et al.* have tested liner-SVM and RBF-SVM methods to predict the conversion of AD by using longitudinal FDG-PET from baseline to 24 months. The results have shown that the linear-SVM can give a better performance. Therefore, each method has its own advantages and no method can yield high-level results for all the tasks and all the time.

In addition, other classifiers are also used in AD diagnosis. Hinrich *et al.* [70] utilized voxel-wise features derived from FDG-PET to predict AD under a boosting framework. Gray *et al.* [38] exploited random forests to derive the pairwise similarity measures from features and then made a classification by using random forests as well for multi-modality, in which FDG-PET was included. Shi *et al.* [75] also focused on multi-modality, and presented a comparative result for distinguishing AD from NC by utilizing a coupled boosting method.

Table 2.3 lists recent studies relevant to FDG-PET in which we focus their performance for two tasks, AD vs. NC and pMCI vs. sMCI. The methods that have not addressed the problem of pMCI prediction but MCI classification (AD vs. MCI or MCI vs. NC) are marked with footnotes. In addition, the list order in Table 2.3 is the same with Table 2.1, and the two tables focus on different aspects of CAD methods.

2.6 Deep learning techniques

In contrast to those conventional machine learning algorithms, Deep learning is a kind of emerging technique and has gained an increasing number of attention

Table 2.3: A brief overview of classifiers used in CAD methods for Alzheimer's disease.

Method	Subjects				Classification	Accuracy(%)	
	AD	NC	pMCI	sMCI	& CV	AD/NC	pMCI/sMCI
Vandenberghe *et al.* [37]	27	25	20[a]		SVM & LOO	100	—
Gray *et al.* [38]	37	35	34	41	RF & Hold-out	89	58
Tong *et al.* [39]	37	25	75[a]		RF & Hold-out	88.6	75.4[b]
Salas *et al.* [69]	53	52	114[a]		SVM & LOO	92	86[b]
Padilla *et al.* [68]	53	52	—	—	SVM & LOO	86.59	—
Hinrichs *et al.* [70]	89	94	—	—	Boosting & 2-fold	84	—
Cabral *et al.* [71]	—	—	44	56	SVM & 10-fold	—	85
Illán *et al.* [72]	95	97	45	164	SVM & 2-fold	88.24	70.21[b]
Pagani *et al.* [49]	—	109	62	—	SVM & LOO	—	91[b]
Li *et al.* [47]	25	30	29[a]		SVM & 10-fold	89.1	64.6[b]
Garali *et al.* [52]	81	61	—	—	SVM & LOO	94.36	—
Garali *et al.* [53]	81	61	29[a]		SVM & LOO	95.07	75.05[b]
Li *et al.* [54]	130	162	130[a]		SVM & Hold-out	91.5	83.1[b]
Gray *et al.* [55]	50	54	53	64	SVM & 4-fold	88.4	63.1
Teipel *et al.* [61]	—	—	39	88	LR & 10-fold	—	72
Zhang *et al.* [73]	51	52	99[a]		SVM & 10-fold	93.2	76.4[b]
Shi *et al.* [75]	51	52	130[a]		Boosting & 10-fold	94.7	—
Zu *et al.* [76]	51	52	43	56	SVM & 10-fold	95.95	69.78
Zhu *et al.* [77]	51	52	43	56	SVM & 10-fold	93.3	69.9

RF = Random Forest
LR = Logistic Regression

[a] subjects of MCI
[b] MCI/NC

due to the impressive performance in recognition and classification tasks [110, 111]. The idea behind this technique is that low-level features can be automatically transformed to high-level features through setting multiple layers. It means the typical feature engineering of classical machine learning methods, including feature design, extraction and selection or reduction, are no longer required, which is paid by huge computational resources. The term 'deep' of deep learning lies in a great number of layers, and such layers can extract abstract features from raw data. Even though numerous deep learning methods have been proposed during the past decade, its architecture can be roughly grouped into three categories, including MultiLayer Perceptron (MLP), Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). Due to the fact that the latter is generally used in processing time series data, such as speech processing and natural language processing, we do not describe it in details in this thesis.

2.6.1 MultiLayer Perceptron

MultiLayer Perceptron (MLP) is a typical deep learning structure and also referred to deep feedforward network. As shown in Figure 2.7, it usually consists of several layers, input layer (yellow), hidden layers (blue) and output layer (red), each layer containing several neurons. The input layer is for receiving inputs, and the hidden layers play the role of feature transformation, while the output layer is used to deliver results. The principle idea of MLP is that each layer in the network is a linear combination of the previous layer outputs combined with a non-linear transformation, which can be expressed as

$$y_i^{l+1} = f(\sum_j W_{ji}^l y_j^l + b_i^l) \tag{2.34}$$

where y_j^l is the output of the j-th neuron in l-th layer, W and b are the weight and bias, respectively, and $f(\cdot)$ is a non-linear transformation, which is also known as the activation function. If the activation is a sigmoid function, the feedforward network is locally considered as the logistic regression. The commonly used activation functions could be ReLU (Rectified Linear Unit), hyperbolic tangent and sigmoid, as shown in Figure 2.8. In addition, the feedforward network is typically trained with using error backpropagation (BP) algorithm [112].

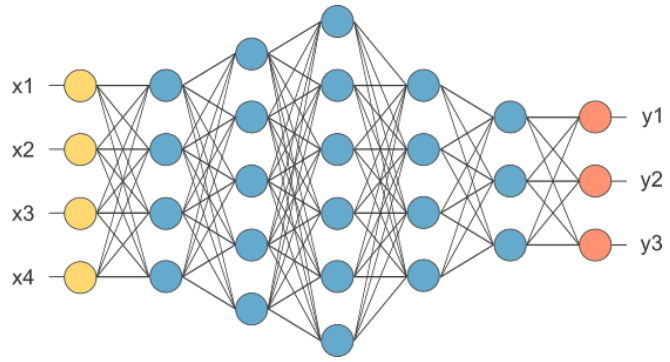


Figure 2.7: An instance of MultiLayer Perceptron

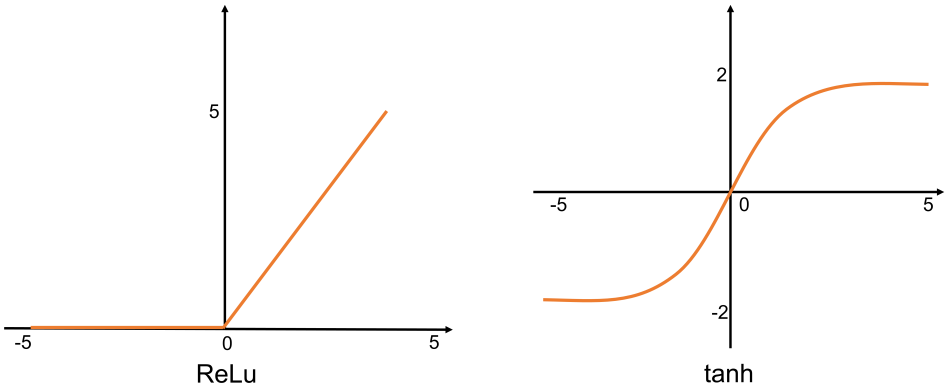


Figure 2.8: An illustration of Relu and tanh functions

2.6.2 Convolutional Neural Network

Convolutional Neural Network (CNN) is suitable for processing spatial data and is widely used in computer vision. The structure of CNN is generally constituted of convolutional layers followed by activation functions, pooling layers and fully-connected layers. The convolutional layer can maintain the spatial continuity of the image and extract the local features. The pooling layer is applied to reduce the dimension of the previous layer output thereby reducing the computing consumption of the next layer and meanwhile provide the rotation invariance. Figure 2.9 shows an instance of the convolution and pooling implementations. Convolution can be considered as the dot product between an input and a kernel function for simplicity sake. For example, in Figure 2.9 colored by yellow, the kernel function is $\begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}$. A 6×6 matrix is then converted to a 4×4 matrix after convolution,

specifically,

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix} = 3, \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix} = 1, \tag{2.35}$$

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix} = 3, \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix} = 2. \tag{2.36}$$

which is corresponding to the first row of the 4×4 matrix. Pooling is easy to understand, either takes the maximum within a window (max pooling) or the mean value (averaging pooling). Before feeding the outputs of the last pooling layer into a fully-connected layer, the outputs should be flatten to a long vector, as shown in Figure 2.10 where the output of the pooling layer is still a matrix. The fully-connected layer plays the role of classifier in CNN. The kernel function consists in several weights, which are parameters to be learned as well as those in the fully-connected layer.

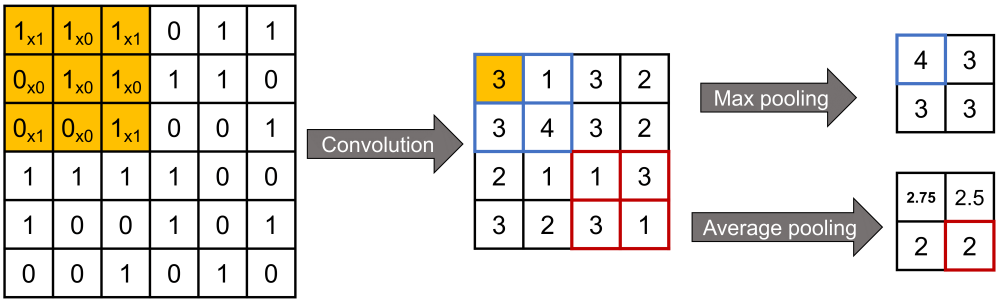


Figure 2.9: An illustration of convolution and pooling

The output size after each operation, either convolution or pooling, is determined by three parameters, the kernel or window size, sliding stride of a window and the padding size,

$$O = \frac{I - K + 2P}{S} + 1 \tag{2.37}$$

where O and I stand for the output and input, respectively, K and P indicate the size of a kernel or padding, respectively, and S is the stride. In addition, the kernel size in convolution step and pooling step can be adjusted according to tasks, and the commonly used size is 3×3, 5×5, 7×7, while for the pooling operation, the window size is usually set to 2×2.

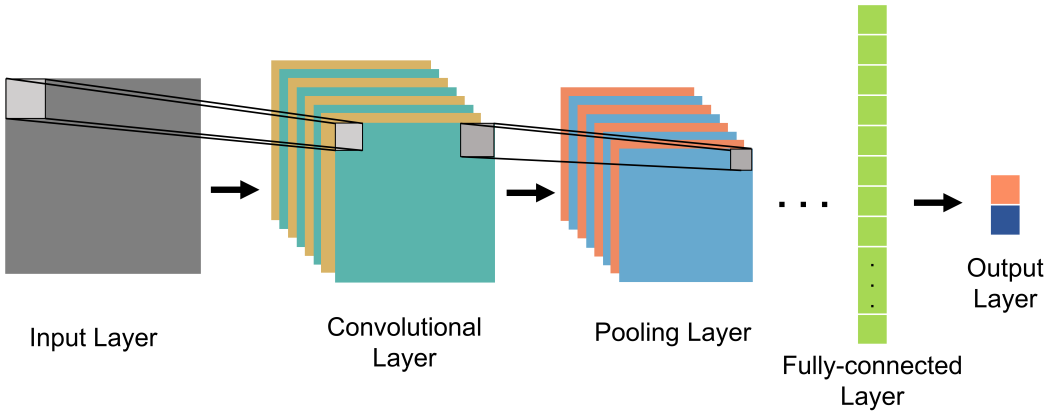


Figure 2.10: An illustration of general CNN architecture.

2.6.3 Applications to neuroimaging

For applying deep learning techniques to neuroimaging data, there could be roughly three groups according to different inputs, which take the ROI/patch value, 2D slice and 3D subject or patch as the input, respectively. Table 2.4 shows different methods using deep learning techniques in neuroimaging-related applications. Due to the limited number of FDG-PET studies, some studies on other modalities have also been considered, such as MRI, Amyloid PET.

ROI/patch value

Methods belonging to this group usually segment subjects into different regions at first, and mean values of regions are taken as the inputs and fed into either MLPs [113] or AutoEncoders (AEs) [114, 115] to achieve the prediction. These methods continue the classical machine learning methods that requires hand-crafted features. But the former just needs very shallow features, since such features can be transformed to abstract features after a series of linear and nonlinear operations, while for the conventional methods, discriminant features are required to design.

2D image

These kinds of methods generally extract 2D slices from 3D neuroimaging data [116, 117] or reconstruct to a 2D image [118], then feed these 2D images to a CNN-based model. Benefit from the success of 2D CNN in natural scene images, this kind of methods can take advantage of the existing CNN models which have been pre-trained in a large dataset and then fine-tune with their local data [118]. It could save a lot of computing resources and give good results. Besides, a 3D image is

decomposed into multiple slices, which can increase the amount of data to some extent. However, since slices are treated independently, some 3D information could be lost.

3D image

Methods belonging to this category either take the whole subject [119–121] or a sub-subject, a 3D patch [122, 123], as the input, thus a 3D CNN is exploited. The main advantage is that the spatial information is fully considered, but more parameters need to be learned, which would be more dependent on high-performance computing resources and the number of dataset. As the boost of computing resources in recent, this kind of approaches have become a trend and are receiving more and more attention.

2.7 Conclusion

We systematically introduced the concepts and algorithms involved in CAD methods in this chapter, from feature extraction, feature selection to classification, as well as the emerging deep learning techniques, which provides a theoretical basis for the following chapters. Moreover, a brief overview of recent research related to AD diagnosis or MCI conversion prediction has been presented.

Table 2.4: A brief overview of deep learning techniques used in CAD methods for Alzheimer's disease.

Method	Modality	Subjects				Input	Accuracy(%)	
		AD	NC	pMCI	sMCI		AD/NC	pMCI/sMCI
Lu *et al.* [114]	FDG-PET	226	304	112	409	Patch value	93.58	82.51
Zhou *et al.* [113]	MRI, FDG-PET, SNP	190	226	389		ROI value	90	74
Liu *et al.* [115]	MRI, FDG-PET	85	77	67	102	ROI value	91.4	82.1
Liu *et al.* [116]	FDG-PET	93	100	146		Slice	91.2	78.9
Ding *et al.* [118]	FDG-PET	484	764	861		2D image	76	55
Gupta *et al.* [117]	MRI	200	232	411		Slice	94.74	86.35
Huang *et al.* [119]	MRI, FDG-PET	647	731	326	441	Subject	90.1	72.22
Spasov *et al.* [120]	MRI, demographic and genetic data	192	184	181	228	Subject	—	86
Hosseini-Asl *et al.* [121]	MRI	70	70	70		Subject	99.3	94.2
Lian *et al.* [122]	MRI	358	429	205	465	3D patch	90.3	80.9
Li *et al.* [123]	MRI	199	229	403		3D patch	89.5	73.8

Chapter 3

ADNI Data

3.1 Introduction

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal multi-center study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease. ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. Four phases are included, which are ADNI 1 (from 2004 to 2010), ADNI GO (from 2009 to 2011), ADNI 2 (from 2011 to 2016) and ongoing ADNI3. Different types of data are available on ADNI, including clinical data, genetics data, MRI, PET and biospecimens, as well as the corresponding longitudinal data. The data is free download for authorized investigators from Image and Data Archive (IDA) of Laboratory of Neuro Imaging (LONI)[1]. In the following of this chapter, we mainly introduce the data acquisition from ADNI, data selection, and the processing procedure.

3.2 Data acquisition

In this thesis, the term 'Data acquisition' does not mean the protocol of getting PET images from the scanner but obtaining from LONI IDA.

3.2.1 FDG-PET in ADNI

FDG-PET images may be different due to different scanners, different shapes and positions of subjects' brains, *etc.* In order to make a consistent starting point for

[1]https://ida.loni.usc.edu/login.jsp?project=ADNI

the subsequent data analysis, ADNI provided four types of pre-processed FDG-PET data, which are 1) Co-registered Dynamic; 2) Co-registered, Averaged; 3) Co-reg, Avg, Standardized Image and Voxel Size; 4) Co-reg, Avg, Std Img and Vox Siz, Uniform Resolution.

Type 1: Co-registered Dynamic Separate frames are extracted from the image file for registration purposes. Either six five-minute frames (ADNI1) or four five-minute frames (ADNI GO/2) are acquired 30 to 60 minutes post-injection. Each extracted frame is co-registered to the first extracted frame of the raw image file (frame acquired at 30-35 min post-injection). The base frame image and the five co-registered frames (or all co-registered frames for the quantitative studies) are recombined into a co-registered dynamic image set. These image sets have the same image size (for example, $128 \times 128 \times 63$ voxels) and voxel dimensions (for example, $2.0 \times 2.0 \times 2.0\ mm^3$) and remain in the same spatial orientation as the original PET image data. This is called **native space**.

Type 2: Co-registered, Averaged This type of processed image set is generated simply by averaging the 6 five-minute frames (or the last 6 frames for the quantitative studies) of co-registered Dynamic image set. This creates a single 30 min PET image set still in **native space**. Type 1 and Type 2 data are only available for PET scans acquired under protocol 1 or 3 (ADNI 1 and ADNI 3).

Type 3: Co-reg, Avg, Standardized Image and Voxel Size Each subject's Type 2 image from their baseline PET scan is then reoriented into a standard $160 \times 160 \times 96$ voxel image grid, having $1.5 \times 1.5 \times 1.5\ mm^3$ voxel size. This image grid is oriented such that the anterior-posterior axis of the subject is parallel to AC-PC line (AC: Anterior Commissure, PC: Posterior Commissure). This is referred to as **AC-PC space**. This standardized image then serves as a reference image for all PET scans on that subject. The individual frames from each PET scan (the baseline study as well as all subsequent studies (6-month scan, 12-month scan, *etc.*) are co-registered to this baseline reference image. By doing the co-registration from the original raw image data to a standardized space in a single step, only one interpolation of the image data is required, and thus resolution degradation by interpolation is kept to a minimum, and is the same for all scans. An averaged image is generated from the AC-PC co-registered frames and then intensity normalized using a subject-specific mask so that the average of voxels within the mask is exactly

one.

Type 4: Co-reg, Avg, Std Img and Vox Siz, Uniform Resolution These images are the result of smoothing of the Type 3 images. Each image set is filtered with a scanner-specific filter function (can be a non-isotropic filter) to produce images of a uniform isotropic resolution of 8 mm full width at half maximum (FWHM), the approximate resolution of the lowest resolution scanners used in ADNI. Image sets from higher resolution scanners have been smoothed more than image sets from lower resolution scanners.

3.2.2 FDG-PET downloaded from IDA

Before downloading data from IDA, users must apply an account and a brief description of their project is needed. Within one week, ADNI will provide an authorization to download and use the data. When you access IDA, click 'SEARCH' and then 'Advanced Image Search (beta)', you will see the searching page. The left column is 'Search Options', including 'SEARCH SECTION' and 'IMAGE TYPES', and the details are shown in Fig. 3.1.

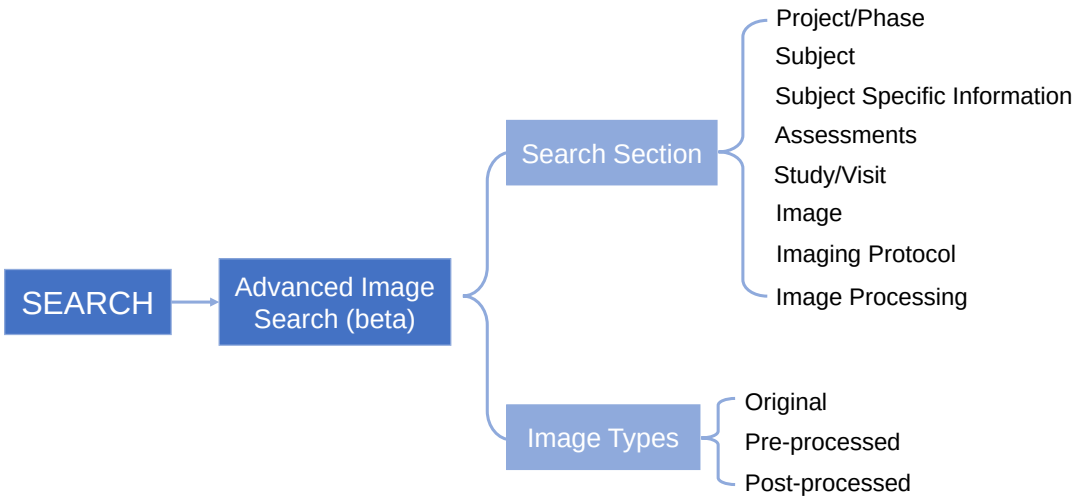


Figure 3.1: Image Searching Options

Project/Phase is the mandatory option and indicates projects of ADNI and AIBL (Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing, not included in this thesis). There are 4 options for project ADNI, including ADNI 1, ADNI GO, ADNI 2 and ADNI3. If you are interested in one or multiple specific phases, then tick their boxes. If you just tick the box of term 'ADNI', the results from all the phases will be included.

Subject contains 5 sub-options, Subject ID, Age (years), Sex, Weight (kgs) and Research Group.

- Subject ID is to specify the subject, and separate multiple IDs by commas if needed.
- Age and Weight are to specify subject's age and weight, you can choose 'Equals' or 'Between' to assign the range.
- Sex, you can choose 'Both', 'Female', 'Male' or 'Unknown'.
- Research Group includes 9 groups, which are MCI, LMCI (Late MCI), EMCI (Early MCI), Patient, AD, Phantom, SMC (Significant Memory Concern, a new cohort added in ADNI 2), Volunteer (individuals who were scanned only as part of the MRI scanner qualification process and did not participate in the study) and CN (Cognitively Normal, same to NC). Only AD, MCI and NC are taken into account in this thesis.

Subject Specific Information is the gene information related to AD.
Assessments contain several clinical tests to measure cognitive impairment.

- FAQ Total Score, Functional Activities Questionnaire;
- GDSCALE Total Score, Geriatric Depression Scale;
- Global CDR, Clinical Dementia Rating Scale;
- MMSE Total Score, Mini-Mental State Examination;
- NPI-Q Total Score, Neuropsychiatric Inventory Questionnaire.

All the measures can be specified by choosing 'Equals' (indicates a specific score) or 'Between' (indicates a range).
Study/visit indicates the 'Study Date', 'Archive Date' and the scan time point. 'ADNI Screening/Baseline' is the initial visit and '...Month X' means X months after the first visit.
Image is about 'Image Description', 'Image ID' and 'Modality'

- Image Description is a brief description of the data and usually displayed in the results.
- Image ID is the ID of each image and different from Subject ID. Different images may be from the same subject.

- Modality includes different modalities, DTI, MRI, PET and fMRI, you should choose at least one modality.

Imaging Protocol is related to imaging parameters, such as Manufacturer, Slice Thickness, Weighting, Radiopharmaceutical *etc.*
Image Processing contains five sub-options, which are 'Image File Type', 'Anatomic Structure', 'Tissue Type', 'Laterality' and 'Registration'. Among them, 'Registration' needs to be paid more attention. This option indicates which space the image is in, Native, AC-PC, Talairach, MNI152, ICBM53 or Colin27.

- Native indicates the image is in the original space.
- AC-PC means the anterior-posterior axis of the image is parallel to AC-PC line.
- Talairach is defined by making two anchors, the anterior commissure and posterior commissure, lie on a straight horizontal line.
- MNI152 is the average of 152 normal MRI scans that have been matched to the MNI305 using a 9 parameter affine transform
- ICBM53 is an average of 53 T1-weighted MRI scans of young healthy adult brains
- Colin27 is that an individual was scanned 27 times, and the scans were co-registered and averaged to create a very high detail MRI dataset of one brain. This average was matched to the MNI305 as well.

MNI305 is 305 normal T1-weighted MRI brains were linearly co-registered (9-params) to 241 brains that had been co-registered (roughly) to the Talairach coordinate system.
Original indicates the raw data without processing.
Pre-processed is the data which has been processed, such as co-registration, averaging, reorientation. There are 4 types of pre-processed data, as introduced in 3.2.1.
Post-processed is the data processed based on one type of pre-processed data by different ADNI groups. For FDG-PET, 3 types of post-processed data are available in ADNI, among which , 2 types data are registered to Talairach space and the other type is spatially normalized to MNI (Montreal Neurological Institute) space.

Figure 3.2 shows the details of different data and Fig. 3.3 displays an instance with visualization for different types. Users can choose different data according to

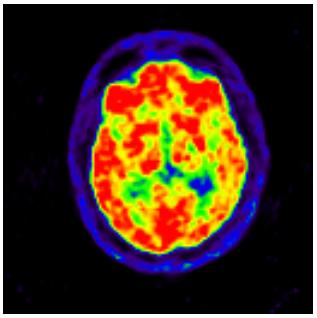
their needs. In this thesis, we use Type 3 FDG-PET data without smoothing to evaluate proposed methods. The options for downloading data are shown in Fig. 3.4.

Image Types	Description	Image Size	Voxel Size (mm^3)	Registration
Original	————————	Uncertain	Uncertain	Native
Pre-processed	Co-registered Dynamic (Type 1)	Uncertain	Uncertain	Native
	Co-registered, Averaged (Type 2)	Uncertain	Uncertain	Native
	Coreg, Avg, Standardized Image and Voxel Size (Type 3)	160x160x96	1.5x1.5x1.5	AC-PC
	Coreg, Avg, Std Img and Vox Siz, Uniform Resolution (Type 4)	160x160x96	1.5x1.5x1.5	AC-PC
Post-processed	Tx, Origin, Spatially Normalized Smoothed	79x95x69	2x2x2	MNI152
	Coreg, Warp, Norm	128x128x60	2.25x2.25x2.25	Talairach
	Talairach Warped	160x160x96	1.5x1.5x1.5	Talairach

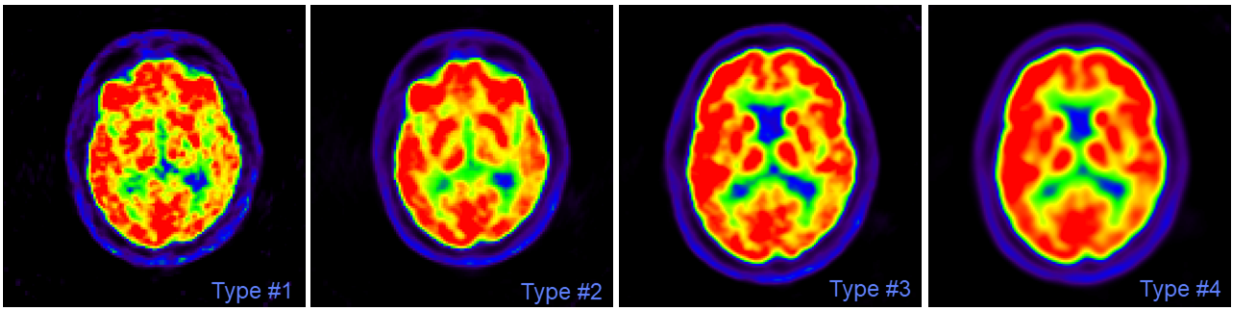
Figure 3.2: Different image types involved in ADNI.

3.3 Data selection

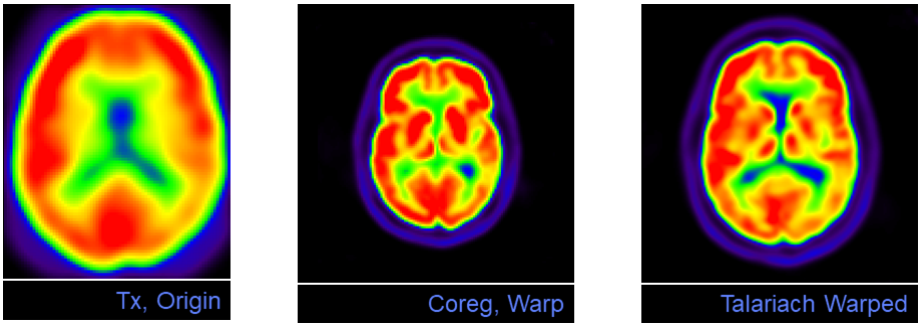
Participants generally take several scans at different time point so as to track their situations. The first time of taking a scan is referred to as the baseline, then 6 months after the baseline, 12 months, 18 months, *etc.* As a result, longitudinal data is available in ADNI dataset. Considering the objective of this thesis is to diagnose AD and predict MCI conversion, only the baseline data is selected. But it is necessary to confirm that the selected data has not changed during its follow-up time. For example, we need to check whether the NC subject has converted to MCI or even AD within a certain month. If the NC subject has converted to MCI during the follow-up period, then the baseline scan from that subject will not be selected. Moreover, as many studies have done, MCI subjects are further classified into two different groups, progressive MCI (pMCI) and stable MCI (sMCI), in order to predict the conversion of an MCI subject. pMCI is the subject who progressed to AD in a certain period of time, while sMCI refers to the subject who keeps stable or reverses to NC. In fact, there is no uniform standard for setting the observation time. Some studies considered the data which does not change within 18 months, some considered data within 24 months or 36 months or even within the available scan time. Clearly, the longer the observation time is, the more challenging the



(a)



(b)



(c)

Figure 3.3: An instance of different image types. (a) Original FDG-PET data. (b) Pre-processed data. (c) Post-processed data

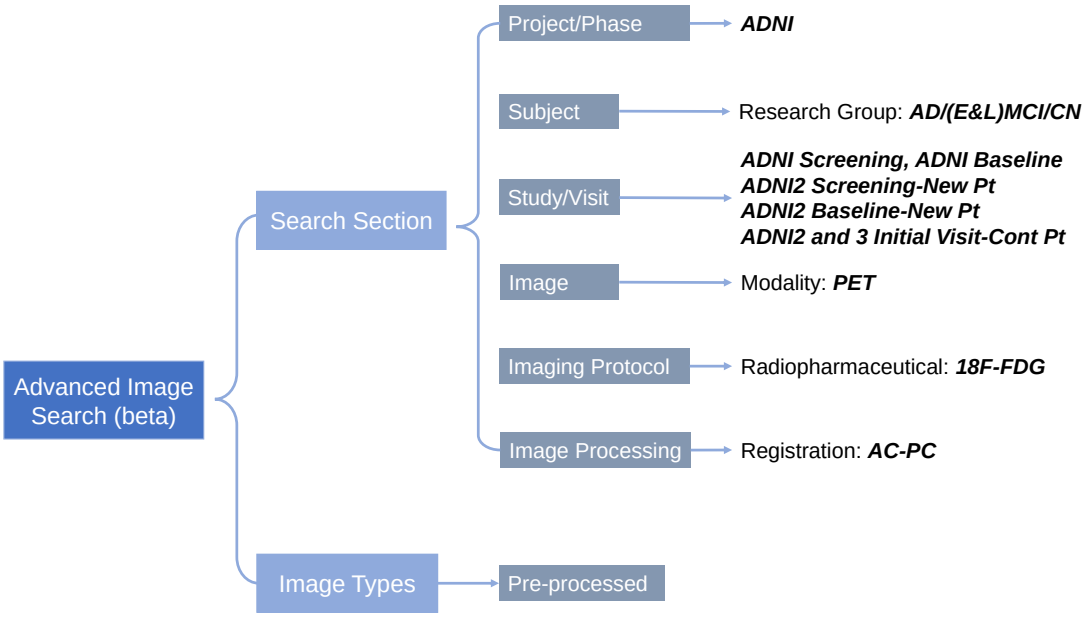


Figure 3.4: FDG-PET Image Searching Options

prediction task. In this thesis, we select the baseline FDG-PET data according to the following rules:

1) AD: subjects diagnosed as AD at the baseline and do not change within the follow-up time;

2) NC: subjects diagnosed as NC at the baseline and do not change within the follow-up time;

3) sMCI: subject diagnosed as MCI at the baseline and stay in the phase of MCI or revert to NC within the available scan time and the visit time is not less than 24 months;

3) pMCI: subjects diagnosed as MCI at the baseline and convert to AD and stay with AD in the available scan time.

ADNI has provided the diagnosis information for all the enrolled subjects, which can be download from IDA through 'Download' $\longrightarrow$ 'Study Data' $\longrightarrow$ 'Assessments' $\longrightarrow$ 'Diagnosis' $\longrightarrow$ 'Diagnostic Summary [ADNI 1,GO,2,3]' $\longrightarrow$ 'DXSUM_PDXCONV_ADNIALL.csv'. The document recorded participant's information in great detail. But for data selection, we just need to focus on 4 variables, 'RID', 'VISCODE2', 'DXCHANGE' and 'DXCURRENT'.

- RID indicates participant roster ID, which is unique. One RID have multiple scans since a subject takes multiple tests at different time points.

- VISCODE2 stands for visiting time points, 'bl' means baseline, 'm06' means 6 months.
- DXCHANGE is for data from ADNI GO and ADNI 2 and means the changing state.

  a) 1 → Stable: NC to NC;
  b) 2 → Stable: MCI to MCI;
  c) 3 → Stable: AD to AD;
  d) 4 → Conversion: NC to MCI;
  e) 5 → Conversion: MCI to AD;
  f) 6 → Conversion: NC to AD;
  g) 7 → Reversion: MCI to NC;
  h) 8 → Reversion: AD to MCI;
  i) 9 → Reversion: AD to NC.

- DXCURRENT is for data from ADNI 1 and indicates the current state.

  a) 1 → NC;
  b) 2 → MCI;
  c) 3 → AD.

Then according to the data selection rules and the diagnosis document, the experimental dataset which consists of 1048 subjects are obtained.

3.4 Data processing

After selecting the data, they are then further processed through the pipeline: reorientation (optional) → spatial normalization → intensity normalization → smoothing. These steps can ensure images in the same standardized space, then the subsequent analysis and comparison make sense. In this thesis, the MNI space is the standardized space for the experimental data. All the procedures are implemented with SPM12 [124], and the processing pipeline is shown in Fig. 3.5.

The reorientation step in this thesis is to ensure the origin of the used image is roughly at AC point as the origin of MNI space is AC, which is an important step for the subsequent spatial normalization. If the image's origin has already pointed

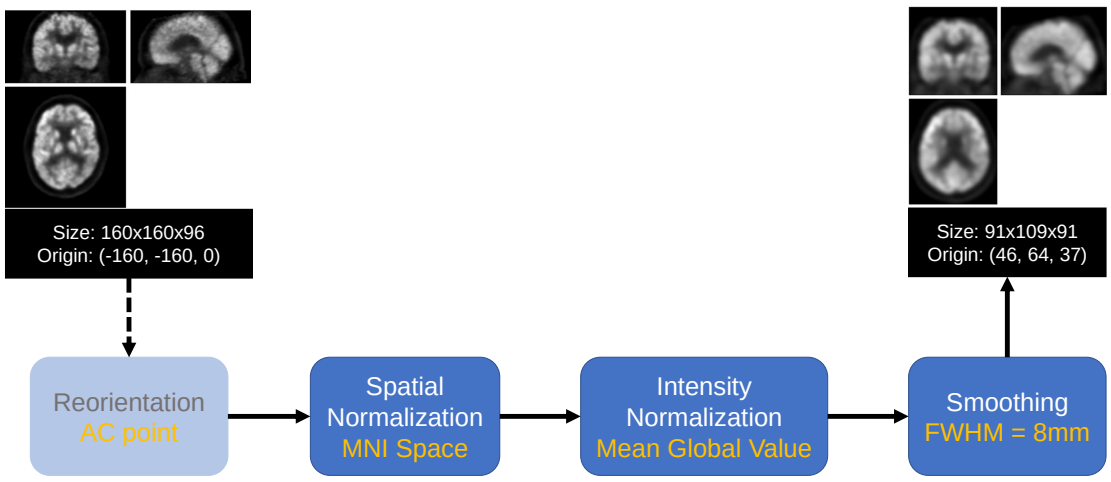


Figure 3.5: FDG-PET Image Processing Pipeline

to AC, then this step is not needed. For the Type 3 image used in this thesis, the origin is at $(-160, -160, 0)$, so the image is manually reoriented to its AC point, approximately at $(80, 89.5, 43.5)$ in its space.

The spatial normalization step is to warp the image into MNI space and make the image have the same size with the template image. Here 'Old Normalize' unit and template image 'PET.nii' embedded in SPM12 are used. All the parameters are defaults except the bonding box, which is reset as $\begin{pmatrix} -90 & -126 & -72 \\ 90 & 90 & 108 \end{pmatrix}$. Through this step, the FDG-PET image is with $2 \times 2 \times 2\ mm^3$ voxel size and $91 \times 109 \times 91$ matrix dimension. In addition, its origin is at $(46, 64, 37)$.

The intensity normalization is then performed through dividing each voxel intensity by the average global value. Thereafter, images are further smoothed by a Gaussian kernel with a full width at half maximum of 8 mm. Through the processing pipeline, 1048 baseline FDG-PET images constitute the experimental data, among which there are 237 subjects with AD, 242 subjects under NC and 569 subjects with MCI, including 209 pMCI and 360 sMCI. The demographic and clinical information of subjects is provided in Table 3.1, in which MMSE stands for the Mini-Mental State Examination.

3.5 Conclusion

In this chapter, we have introduced ADNI dataset, including data acquisition, selection and processing, in order to provide an insight into the experimental data. Consider the objective of this thesis, only baseline data is chosen, totally 1048 im-

Table 3.1: Demographic and clinical information of subjects.

Characteristic	AD	NC	pMCI	sMCI
Number of subjects	237	242	209	360
Female/male	97/140	122/120	87/122	153/207
Age(Mean ± SD)	75.00 ± 7.91	73.66 ± 5.66	73.89 ± 6.88	71.73 ± 7.66
MMSE(Mean ± SD)	23.19 ± 2.12	29.03 ± 1.20	27.13 ± 1.71	28.20 ± 1.59

ages. The subsequent experimental results and analysis are based on this dataset.

Chapter 4

Multilevel Feature Representation for FDG-PET Images

4.1 Introduction

There is a variety of literature applying voxel-wise or ROI-wise (Region of Interest) features to address the problem of AD diagnosis, but connectivities between regions are rarely taken into account. In fact, a human brain is a complex system and multiple regions interact with each other. Therefore, connectivities between regions are important in AD classification or MCI conversion prediction and cannot be ignored. In this chapter, we develop a novel method by using single modality, FDG-PET, but multilevel feature, which considers both region's properties and connectivities between regions to classify AD or pMCI from NC or sMCI, respectively. In the following parts within this chapter, the proposed method is described in details at first and then a series of experiments are conducted to evaluate its performance and validate its effectiveness from different views. Lastly, a conclusion is given.

4.2 Method

The proposed method is described from 3 aspects, including feature extraction, feature selection and ensemble classification, as shown in Figure 4.1. First, after segmenting each subject into 90 ROIs according to an AAL atlas, 3 levels of features are extracted, specifically, the 1st-Level feature, which comprises ROI's mean intensity and standard deviation. The 2nd-Level feature, the similarity-based

connectivity between any pair of ROIs, is decomposed into 3 sets according to a proposed similarity-driven ranking method. The 3rd-Level feature is composed of graph-based features. Next, LASSO is applied to do the feature selection for each set of features, respectively. Then different classifiers are trained using different sets of features. Final prediction is obtained through an ensemble classifier decided by a proposed maximum Mean squared Error (mMsE) strategy and a nested cross validation technique.

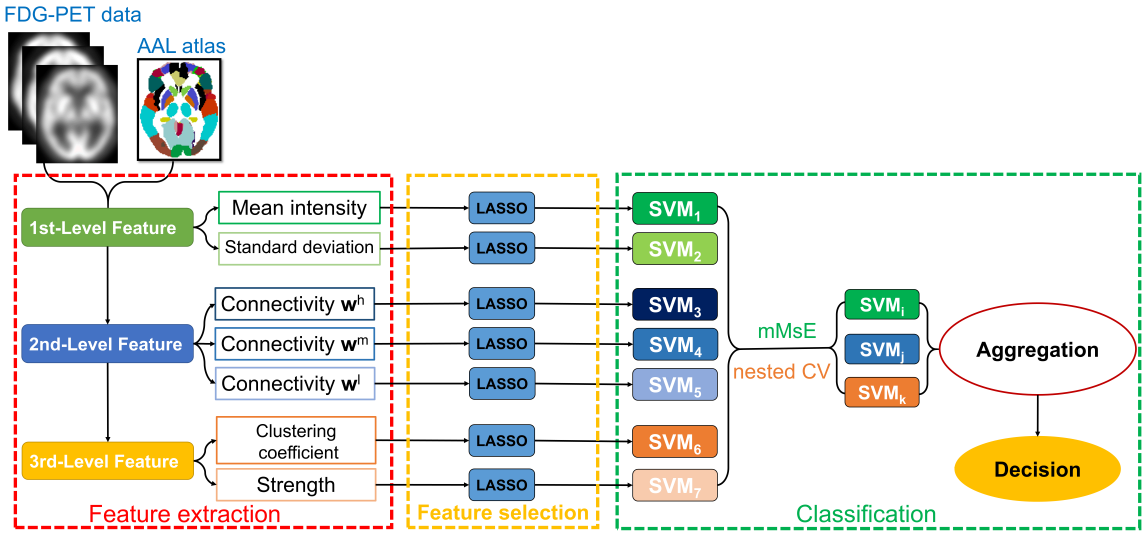


Figure 4.1: The framework of the proposed method.

4.2.1 Feature extraction

Many methods in the existing literature used mean gray level intensities of some ROIs as features [47, 49, 52–55]. However, only ROI's information is not enough. Therefore, in this chapter, we explore to expand the feature pool computed on FDG-PET data.

1st-Level Feature Since each region's mean intensity and standard deviation can reflect the FDG uptake and its corresponding distribution, the 1st-Level feature for the n-th subject can be represented as:

$$\mathbf{r}_n^m = [r_{n1}^m, r_{n2}^m, \cdots, r_{np}^m] \tag{4.1}$$

$$\mathbf{r}_n^s = [r_{n1}^s, r_{n2}^s, \cdots, r_{np}^s] \tag{4.2}$$

where $\mathbf{r}_n^m$ and $\mathbf{r}_n^s$ are the mean intensity and standard deviation, respectively, and p is the number of ROIs, here $p = 90$.

2nd-Level Feature The 2nd-Level feature is the similarity-based connectivity between ROIs. Hereafter, connectivity is used to refer to similarity-based connectivity. First, the 1st-Level feature is used to represent each ROI, and the i-th ROI for a certain subject is represented by:

$$\mathbf{x}_i = [r_i^m, r_i^s] \tag{4.3}$$

then the connectivity between any two ROIs is computed through:

$$w_{ij} = \begin{cases} e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2} & i \neq j, \\ 0 & i = j. \end{cases} \tag{4.4}$$

where w_{ij} is the connectivity of the i-th ROI and the j-th ROI, and the higher the value of w_{ij}, the more similar the two ROIs. It should be noted that before computing w_{ij} through (4.4), each type of the 1st-Level feature is normalized over ROIs. The 2nd-Level feature of any subject is:

$$\mathbf{W_r} = \begin{bmatrix} 0 & w_{r12} & \cdots & w_{r1j} & \cdots & w_{r1p} \\ w_{r21} & 0 & \cdots & w_{r2j} & \cdots & w_{r2p} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ w_{ri1} & w_{ri2} & \cdots & 0 & \cdots & w_{rip} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ w_{rp,1} & w_{rp2} & \cdots & w_{rpj} & \cdots & 0 \end{bmatrix} \tag{4.5}$$

where $\mathbf{W_r}$ is a symmetric matrix. The 2nd-Level feature is composed of connectivities between all the 90 ROIs, totally 4005 dimensions ($90 \times (90-1)/2$, only considering the values on the upper triangle). Clearly, it is not an optimal dimension for the subsequent classification. Therefore, $\mathbf{W_r}$ is further decomposed into 3 subsets of features according to a proposed similarity-driven ranking method.

Similar to the way of computing connectivities between ROIs, we can obtain the similarity coefficients between subjects for a specific ROI:

$$w_{uv} = \begin{cases} e^{-\|\mathbf{x}_u - \mathbf{x}_v\|^2} & u \neq v, \\ 0 & u = v. \end{cases} \tag{4.6}$$

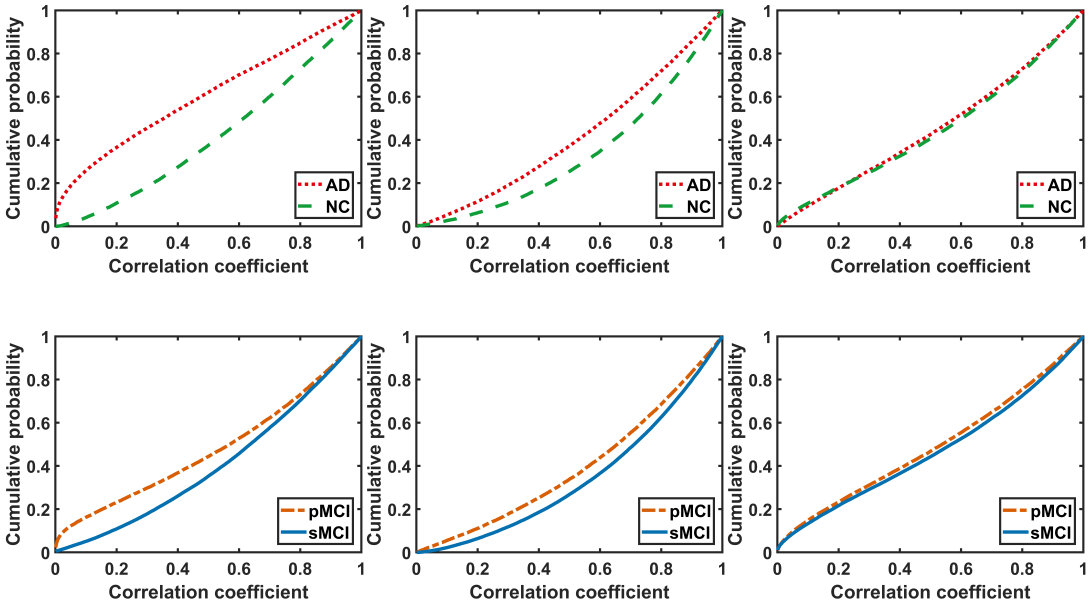
where u, v stands for the u-th and v-th subjects. For any ROI, a symmetric matrix for subjects, $\mathbf{W_s}$, is obtained from:

$$\mathbf{W_s} = \begin{bmatrix} 0 & w_{s12} & \cdots & w_{s1v} & \cdots & w_{s1N} \\ w_{s21} & 0 & \cdots & w_{s2v} & \cdots & w_{s2N} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ w_{su1} & w_{su2} & \cdots & 0 & \cdots & w_{suN} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ w_{sN1} & w_{sN2} & \cdots & w_{sNv} & \cdots & 0 \end{bmatrix} \tag{4.7}$$

The dimension of $\mathbf{W_s}$ is determined by the number of subjects, N, in a group (AD, NC, MCI, pMCI and sMCI). For example, there are 237 subjects in AD group, so $N = 237$, then the dimension of $\mathbf{W_s}$ is 237×237. Each subject is segmented into 90 ROIs, thus there are 90 matrices like $\mathbf{W_s}$.

If taking NC subjects (including training and testing samples) as a reference, in one hand, for a ROI which is not affected by AD, the similarity coefficients between AD subjects are supposed to be close to those of NC subjects. In the other hand, for a ROI affected by AD, the similarity coefficients of AD subjects are different from NC group. On order to quantify the difference, we first make a statistic on the upper triangle values of $\mathbf{W_s}$ to get the frequency distribution histogram of those values. Then the cumulative probability curve of similarity coefficients can be obtained, as shown in Figure 4.2, where (a), (b) and (c) stand for region Angular_L, region Temporal_Sup_R and region Heschl_R, respectively and the top row is for AD vs. NC, while the bottom row is for pMCI vs. sMCI. All the figures share the same x axis. It can be seen that there is a clear difference between the AD and NC groups in Figure 4.2(a). Even though the difference between pMCI and sMCI groups is not as great as that in AD vs. NC, it is reasonable since identifying pMCI from sMCI is more challenging than AD classification. For the other two ROIs, the difference decreases gradually. It implies that among the experimental subjects, region Heschl_R is almost unaffected by AD, while region Angular_L has a great chance of getting influenced, therefore region Angular_L is ranked before region Heschl_R, and region Temporal_Sup_R is placed between them. The difference between curves is computed through the difference of area under curve, which is denoted ΔS. The larger the ΔS, the greater the impact generated by AD for a ROI. At last, all the ROIs can be ranked according to ΔS from high to low, as illustrated in Figure 4.3, from (a) to (b). It should be noted that we highly recommend using a

balance number of subjects in 2 groups for the comparison and the more the better.



(a) ROI: Angular_L (b) ROI: Temporal_Sup_R (c) ROI: Heschl_R

Figure 4.2: Statistics of the similarity coefficients between subjects for a certain ROI. Top row: AD vs. NC. Bottom row: pMCI vs. sMCI.

After ranking all the ROIs according to the proposed method, the similarity matrix $\mathbf{W_r}$ is re-calculated according to the new order of ROIs. Then $\mathbf{W_r}$ is divided into 4 equal parts, as shown in Figure 4.3(b), where the red part stands for the set of regions which are highly influenced by AD, denoted $\mathbf{W}^h$, while the blue part stands for ROIs with less impact of AD, denoted $\mathbf{W}^l$, and the green part represents the connectivities between highly influenced ROIs and slightly influenced ROIs, which is denoted $\mathbf{W}^m$. Since $\mathbf{W_r}$ is symmetric, only upper triangular matrix is taken into consideration, like in Figure 4.3(c). Therefore, the 2nd-Level feature $\mathbf{W_r}$ is divided into 3 sets, and after converting them to vectors, the 2nd-Level feature for the n-th subject is represented as:

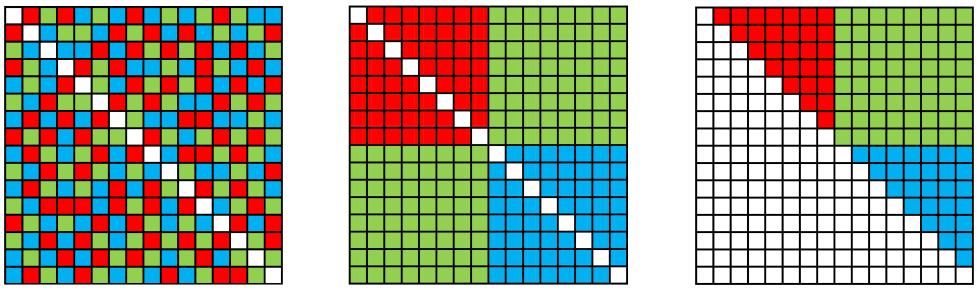
$$\mathbf{w}_n^h = [w_{n1}^h, w_{n2}^h, \cdots, w_{np^h}^h] \tag{4.8}$$

$$\mathbf{w}_n^m = [w_{n1}^m, w_{n2}^m, \cdots, w_{np^m}^h] \tag{4.9}$$

$$\mathbf{w}_n^l = [w_{n1}^l, w_{n2}^l, \cdots, w_{np^l}^l] \tag{4.10}$$

where p^h, p^m and p^l are the dimension of each subset of features. p^h and p^l are the same (red and blue parts in Figure 4.3(b)), both equal to 990 ($45 \times (45-1)/2$), and p^m (green part) is 2025 (45×45). Apparently, compared to 4005 (red, blue

and green parts), the dimension is decreased by about 50%–75%. Table 4.1 and Table 4.2 list the top 20 regions which are highly and slightly relevant to AD or pMCI, respectively.



(a) (b) (c)

Figure 4.3: Instance of the division for a similarity matrix $\mathbf{W_r}$.

3rd-Level Feature The 3rd-Level feature is extracted from a graph point of view, which stands for an overall connectivity between a ROI and the other ones. Generally, a graph $G = (V, E)$ consists of a finite set V of vertices and a finite set of edges $E \subseteq V \times V$. A vertex in a graph is equivalent to a ROI in a brain. Therefore, the connectivity between the i-th ROI and the j-th ROI, w_{ij}, can be viewed as the weight of an edge which connects the i-th vertex and the j-th vertex. In this chapter, we analyze the undirected graph, which means $w_{ij} = w_{ji}$. Then a subject can be represented by a graph, as shown in Figure 4.4 [125] which represents a subject from ADNI database.

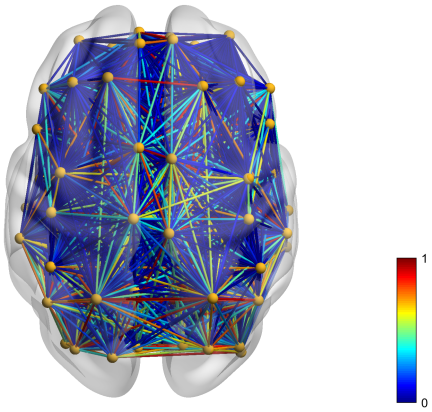


Figure 4.4: Instance of the brain connectivity network from the axial view.

After constructing a graph for a subject, several graph measures can be computed, such as degree, strength, clustering coefficient, betweenness centrality [126].

Table 4.1: Top 20 ROIs highly relevant to AD or pMCI

Rank	AD vs. NC	pMCI vs. sMCI
1	Fusiform_R	Fusiform_R
2	Temporal_Inf_L	Precuneus_R
3	Fusiform_L	Fusiform_L
4	Angular_R	Occipital_Mid_R
5	Occipital_Mid_R	Temporal_Inf_L
6	Temporal_Mid_R	Olfactory_R
7	Angular_L	Temporal_Inf_R
8	Cingulum_Post_L	Temporal_Sup_L
9	Cingulum_Post_R	Angular_R
10	Occipital_Mid_L	Temporal_Mid_R
11	Temporal_Inf_R	Parietal_Sup_R
12	Temporal_Mid_L	Parietal_Sup_L
13	Precuneus_R	Occipital_Inf_R
14	Parietal_Inf_L	Precuneus_L
15	Occipital_Sup_R	Hippocampus_L
16	Parietal_Inf_R	Angular_R
17	Occipital_Inf_R	Rectus_L
18	Hippocampus_R	Occipital_Sup_L
19	Cuneus_R	Frontal_Mid_Orb_R
20	Hippocampus_L	Occipital_Sup_R

Table 4.2: Top 20 ROIs slightly relevant to AD or pMCI, Rank -1 indicates the most irrelevant to AD or pMCI

Rank	AD vs. NC	pMCI vs. sMCI
-1	Heschl_R	Frontal_Sup_Medial_R
-2	Supp_Motor_Area_L	Cingulum_Ant_L
-3	Thalamus_L	Thalamus_L
-4	Calcarine_L	Frontal_Sup_Medial_L
-5	Calcarine_R	Heschl_L
-6	Heschl_L	Frontal_Inf_Oper_L
-7	Frontal_Sup_Medial_L	Calcarine_L
-8	Rolandic_Oper_R	Heschl_R
-9	Rolandic_Oper_L	Caudate_L
-10	Frontal_Inf_Oper_L	Calcarine_R
-11	Cingulum_Ant_L	Rolandic_Oper_R
-12	Frontal_Sup_R	Rolandic_Oper_L
-13	Supp_Motor_Area_R	Temporal_Pole_Sup_R
-14	Precentral_R	Putamen_R
-15	Cuneus_L	Cingulum_Ant_R
-16	Frontal_Sup_Orb_R	Frontal_Inf_Orb_L
-17	Temporal_Pole_Sup_R	Temporal_Pole_Sup_L
-18	Precentral_L	Postcentral_L
-19	Frontal_Sup_Medial_R	Frontal_Sup_R
-20	Thalamus_R	Frontal_Sup_L

According to [127, 128], the metrics strength and clustering coefficient are effective in discriminating AD, therefore the 3rd-Level feature is represented by these two graph measures. Specifically,

strength: the sum of a vertex's neighboring link weights [126].

$$s_i = \sum_{j=1}^{p} w_{ij} \tag{4.11}$$

where s_i is the strength of a vertex or a ROI.

clustering coefficient: the geometric mean of all triangles associated with each vertex [126].

$$\mathbf{c} = \frac{diag((\mathbf{W_r} \cdot^{\frac{1}{3}})^3)}{\mathbf{d}(\mathbf{d}-1)} \tag{4.12}$$

where $diag(\cdot)$ is an operator which takes the diagonal values from a matrix, $\mathbf{c}$ is a clustering coefficient vector, and $\mathbf{d}$ is a degree vector in which the element d_i is,

$$d_i = \sum_{j=1}^{p} a_{ij} \tag{4.13}$$

where a_{ij} is the connection status between the i-th vertex and the j-th vertex: $a_{ij} = 0$ when $w_{ij} = 0$, otherwise $a_{ij} = 1$.

Thus, the 3rd-Level feature consists of 2 sets of features, and each of them for the n-th sample is represented as:

$$\mathbf{g}_n^s = [s_{n1}, s_{n2}, \cdots, s_{np}] \tag{4.14}$$

$$\mathbf{g}_n^c = \mathbf{c}_n \tag{4.15}$$

These features exhibit different ranges of values. Thus a procedure of feature normalization is necessary by z-score prior to classification:

$$z_{nt} = \frac{f_{nt} - \mu_t}{\delta_t} \tag{4.16}$$

where f_{nt} is the value of the t-th feature of the n-th subject, and $f \in \{r^m, r^s, w^h, w^m, w^l, g^s, g^c\}$, μ_t and δ_t are the mean value and standard deviation of the t-th feature, respectively. Most of f_{nt} values can be transformed to the range $[-1, 1]$ through (4.16), while out-of-range values are clamped to either -1 or 1.

4.2.2 Feature selection

We have proposed 3 levels of features in our method. For the 1st-Level and 3rd-Level features, the dimension is 90 for each type of feature. For the 3 subsets of features in 2nd-Level, the dimension is 990 ($\mathbf{w}^h$), 2025 ($\mathbf{w}^m$), 990 ($\mathbf{w}^l$), respectively. Therefore, it is necessary to select representative features to reduce the feature dimension. A good strategy of feature reduction or selection is to remove irrelevant, redundant and noisy features and meanwhile improve classification performances. Least Absolute Shrinkage and Selection Operator (LASSO) is one of the popular techniques for dimension reduction and feature selection. It uses l_1 regularization to get a sparsity solution, thereby achieving the goal of feature selection. In this work, feature selection is accomplished by using LASSO, which has been introduced in Chapter 2.

4.2.3 Ensemble classification

The support vector machine (SVM) classifier is a popular and effective method in distinguishing subjects with AD or MCI from NC. In this study, 3 levels of features, which then are decomposed into 7 types of features, are fed into 7 linear SVMs to train 7 individual models, respectively. The motivation of training in this way is to ensure a model focus on one type of feature of the data. The margin parameter C of all the SVMs is fixed to 1 for a fair comparison.

The effectiveness of an ensemble classifier depends on the number of individual classifiers and the diversity between them. The more the number of classifiers and the higher the diversity, the more effective the ensemble classifier is. However, if the sub-classifier doesn't perform well (the accuracy is usually between 50% and 60%), the increase of the number of classifiers cannot improve the ensemble classifier's performance, because as the number of classifiers increases, the possibility that misclassified results accounted for the majority also increases. Thus, in order to enhance the ensemble effect and meanwhile, avoid misclassified results taken up the majority, a strategy of selecting models, maximum Mean square Error (mMsE), is proposed. Let $\mathbf{y}_i$ and $\mathbf{y}_j$ denote the output labels of SVM$_i$ and SVM$_j$, respectively, then the Mean Square Error (MSE) between $\mathbf{y}_i$ and $\mathbf{y}_j$ is computed through,

$$M(i,j) = \frac{1}{K}\|\mathbf{y}_i - \mathbf{y}_j\|^2 \tag{4.17}$$

where K is the number of the testing samples and each element in $\mathbf{y}_i$ belongs to

$\{-1, 1\}$. The higher the MSE, the greater the diversity between the outputs of classifiers. Then a pair of classifiers with high diversity can be achieved by finding the maximum MSE,

$$(i, j) = \arg\max_{i,j} M(i, j) \tag{4.18}$$

In addition, another classifier, $\mathbf{y}_k$, is determined through nested cross validation on the training set and the one with the highest accuracy is selected. Last, the final decision is made through a majority voting of the 3 selected classifiers' outputs:

$$\mathbf{Y} = \text{sgn}(\mathbf{y}_i + \mathbf{y}_j + \mathbf{y}_k) \tag{4.19}$$

where $\text{sgn}(\cdot)$ is a sign function. Even though the number of classifiers for decision making decreases, the classifiers with high diversity and high accuracy are kept. Therefore, the strategy can enhance the ensemble effect, especially in the case where all the classifiers do not have a good performance, since it can avoid misclassified results accounted for the majority.

4.3 Experiments and results

4.3.1 Setup

Experiments are conducted on 2 different kinds of classifications, including AD vs. NC and pMCI vs. sMCI. In order to evaluate the performance of the proposed method, 4 commonly used metrics, classification accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under curve (AUC) that have been introduced in Chapter 2 are used. In addition, balanced ACC (bACC) is applied for the case that positive and negative data are unbalanced, for example, pMCI vs. sMCI in the thesis. The higher the values are, the better the corresponding method is. Because of a limited number of samples, we use a 10-fold cross validation technique to assess the performance, and repeat 10 times to reduce the possible bias. The parameter in LASSO, λ, controls the number of selected featutes and is decided by nested cross validation on the training dataset within the range $\{10^{-5}, 10^{-4}, ..., 10^{-1}\}$ for the 1st-Level and 3rd-Level features, and $\{10^{-10}, 10^{-8}, ..., 10^{-1}\}$ for the 2nd-Level feature. It should be noted that only the results involved in ensemble classifications are obtained by performing LASSO prior to classification. The whole procedure is shown in Algorithm 1. The SVM algorithm is implemented with the LIBSVM toolbox [129].

Algorithm 1 Workflow of the proposed method.

1: Dividing the dataset into 10 parts, one of them is used as testing data and the remaining parts are for training;
2: Extracting 7 types of features for the training and testing data, respectively;
3: Selecting features by LASSO for each type of features;
4: Training different models using different types of features on training data;
5: Using the proposed mMsE method and the nested cross validation technique to choose 3 models;
6: Applying the 3 models on testing data and then the evaluation metrics (ACC, SEN, SPE, AUC, bACC) can be computed;
7: Returning to step 1, choosing another part as the testing data till all the 10 parts are used for testing;
8: Repeating step 1 to step 7 ten times, then computing the average value of each metric.

4.3.2 Single-type feature representation evaluation

The 3 levels of features are decomposed to 7 different types of features, and the performance of each type of feature is shown in Table 4.3 and Table 4.4 for AD vs. NC and pMCI vs. sMCI, respectively. It can be seen that region's mean intensity is a kind of effective feature in both of the tasks. Even though it does not achieve the best performance in AD vs. NC, it still yields comparative results with slight differences from the best one ($\mathbf{w}^h$), which are 0.65%, 0.33% and 1.05% in terms of ACC, SEN and SPE, respectively. The connectivity $\mathbf{w}^l$, which stands for connections between ROIs with less impact of AD, is inferior among all the types of features in AD diagnosis and classifying pMCI from sMCI. In addition, the graph features, strength and clustering coefficient, are more useful in pMCI vs. sMCI than they are in AD classification, while standard deviation works better in AD vs. NC. The 3 sets of connectivities have different performance in different tasks. For AD vs. NC, the connectivity $\mathbf{w}^h$ outperforms the other 2 sets, while for identifying pMCI from sMCI, the relative best results come from the connectivity $\mathbf{w}^m$ (bACC: 66.84%) which has a larger dimension than $\mathbf{w}^h$ (bACC: 63.36%). It is due to a trade-off between informative and redundant features. More features can provide more information, which can enhance the classifier's performance. On the other hand, more features can lead to feature redundancy with a high possibility, which could harm the classifier's training.

Table 4.3: Performance of different types of feature for AD vs. NC(%)

Method	Feature	ACC	SEN	SPE	AUC
1st-Level	Mean intensity	88.56	87.08	90.00	**94.91**
1st-Level	Standard deviation	86.94	86.47	87.49	94.77
2nd-Level	Connectivity $\mathbf{w}^h$	**89.21**	**87.41**	**91.05**	94.59
2nd-Level	Connectivity $\mathbf{w}^m$	87.73	86.49	89.05	94.12
2nd-Level	Connectivity $\mathbf{w}^l$	79.42	78.76	80.10	87.31
3rd-Level	Strength	84.86	84.26	85.54	91.53
3rd-Level	Clustering coefficient	85.03	84.80	85.34	91.40

Table 4.4: Performance of different types of feature for pMCI vs. sMCI(%)

Method	Feature	ACC	SEN	SPE	AUC	bACC
1st-Level	Mean intensity	**71.46**	**66.71**	74.42	**77.59**	**70.57**
1st-Level	Standard deviation	65.24	60.20	68.08	71.00	64.14
2nd-Level	Connectivity $\mathbf{w}^h$	66.06	52.79	73.93	67.83	63.36
2nd-Level	Connectivity $\mathbf{w}^m$	69.77	55.73	**77.95**	71.28	66.84
2nd-Level	Connectivity $\mathbf{w}^l$	62.14	46.91	71.10	62.77	59.01
3rd-Level	Strength	69.76	65.89	72.12	75.71	69.01
3rd-Level	Clustering coefficient	68.14	64.64	70.37	74.97	67.51

4.3.3 Feature selection evaluation

In order to minimize the influence caused by feature redundancy and then test the effectiveness of different types of features, LASSO is performed before feeding features to SVM. The best results with feature selection are shown in Table 4.5 and Table 4.6, and the values in parentheses indicate increases or decreases compared to features without feature selection. It can be seen that the feature selection strategy can affect the performance for all the types of features in different tasks. Most of them are improvements, especially for $\mathbf{w}^h$ in pMCI vs. sMCI. The corresponding improvements regarding ACC, SEN, SPE, AUC and bACC are 7.12%, 11.60%, 4.43%, 10.68% and 8.02% respectively, which are significant. Meanwhile, $\mathbf{w}^h$ also achieves the highest improvement in classifying pMCI from sMCI where SEN is increased by 11.60%, which implies connectivity $\mathbf{w}^h$ is a kind of potential feature. In addition, the mean intensity still has dominant overall performance in the two kinds of classifications, and connectivity $\mathbf{w}^l$ is still not that effective among all the

features. Furthermore, after feature selection, the connectivity from highly effective to relatively effective is $\mathbf{w}^h$, $\mathbf{w}^m$ and $\mathbf{w}^l$ for both AD vs. NC and pMCI vs. sMCI, which is consistent with their properties.

Table 4.5: Performance of feature selection for AD vs. NC(%)

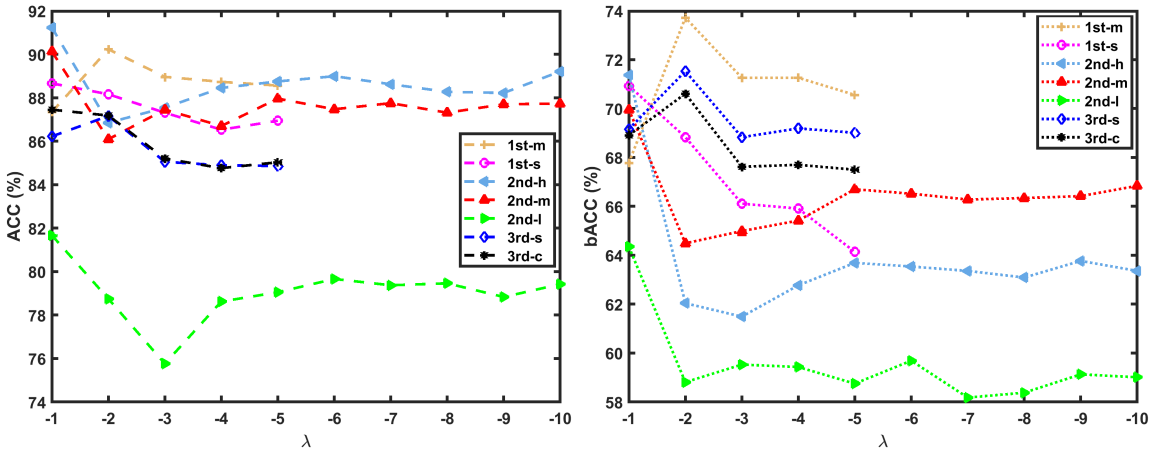
Feature	ACC	SEN	SPE	AUC
Mean intensity	90.23(+1.67)	89.38(+2.30)	91.16(+1.16)	**96.56**(+1.65)
Standard deviation	88.66(+1.72)	86.07(−0.40)	91.28(+3.79)	95.17(+0.40)
Connectivity $\mathbf{w}^h$	**91.23**(+2.02)	**89.39**(+1.98)	**93.01**(+1.96)	96.25(+1.66)
Connectivity $\mathbf{w}^m$	90.13(+2.40)	88.72(+2.23)	91.46(+2.41)	95.86(+1.74)
Connectivity $\mathbf{w}^l$	81.67(+2.25)	77.83(−0.93)	85.45(+5.35)	89.64(+2.33)
Strength	86.22(+1.36)	85.14(+0.88)	87.35(+1.81)	93.55(+2.02)
Clustering coefficient	87.45(+2.42)	85.64(+0.84)	89.35(+4.01)	94.33(+2.93)

Table 4.6: Performance of feature selection for pMCI vs. sMCI(%)

Feature	ACC	SEN	SPE	AUC	bACC
Mean intensity	**74.32**(+2.86)	**71.14**(+4.43)	76.29(+1.87)	**80.49**(+2.90)	**73.72**(+3.15)
Standard deviation	71.21(+5.97)	69.32(+9.12)	72.53(+4.45)	76.46(+5.46)	70.93(+6.79)
Connectivity $\mathbf{w}^h$	73.18(+7.12)	64.39(+11.60)	**78.36**(+4.43)	78.51(+10.68)	71.38(+8.02)
Connectivity $\mathbf{w}^m$	71.39(+1.62)	64.19(+8.46)	75.74(−2.21)	77.60(+6.32)	69.97(+3.13)
Connectivity $\mathbf{w}^l$	66.47(+4.33)	56.20(+9.29)	72.49(+1.39)	71.11(+8.34)	64.35(+5.34)
Strength	72.30(+2.54)	67.60(+1.71)	75.45(+3.33)	78.39(+2.68)	71.53(+2.52)
Clustering coefficient	71.41(+3.27)	66.82(+2.18)	74.40(+4.03)	77.04(+2.07)	70.61(+3.10)

Figure 4.5 illustrates accuracies and balanced accuracies under different numbers of features which are controlled by the parameter λ for AD vs. NC and pMCI vs. sMCI respectively, where '2nd' denotes the original 2nd-Level feature, '2nd-h', '2nd-m' and '2nd-l' denote the decomposed 3 subsets of features and '1st-m', '1st-s', '3rd-c' and '3rd-s' denote the mean intensity, standard deviation (1st-Level feature), clustering coefficient and strength (3rd-Level feature), respectively. Noting that a small λ indicates less features are selected and besides, the range of λ in 1st-Level

and 3rd-Level features is different with that of 2n-Level feature, which is within the range of $[-1, -10]$. It is because that the dimensions of 3 subsets of 2nd-Level feature are more than that of 1st-Level and 3rd-Level features, which means extra values of λ are needed so as to the whole features are considered. It can be seen that less features can achieve significant performance for both tasks, which also proves that the feature selection strategy can strengthen the model performance.



(a) (b)

Figure 4.5: Performance of feature selection under different values of λ and a smaller λ implies less features are selected. (a) AD vs. NC. (b) pMCI vs. sMCI.

4.3.4 Feature concatenation evaluation

In this part, the evaluation for different levels of features is given. Different types of features within the same level are concatenated to a long vector and the results are shown in Table 4.7 and Table 4.8 (the first 3 lines). As can be seen, among all the 3 levels of features, the 2nd-Level feature is superior to other features in two tasks. Even though it cannot achieve the best results for all the five metrics in identifying pMCI from sMCI, the 2nd-Level feature has comparative overall performance considering the highest AUC metric it achieves and slight difference with the best one in terms of bACC, which is only 0.3%. In addition, it can be seen from Table 4.3 and Table 4.7 (AD diagnosis) that concatenation of different types of features cannot improve the performance of AD classification and their metrics are lower than the results obtained using the optimal sub-feature in each level. For pMCI vs. sMCI, as can be seen from Table 4.4 and Table 4.8, only concatenation of 2nd-Level features can yield some improvements, and increase by about 0.67% (ACC), 0.09% (SEN), 1.24% (SPE), 3.01% (AUC) and 0.67% (bACC).

Moreover, the performances of concatenating all the 3 levels of features are also shown in Table 4.7 and Table 4.8 (the last line). It can be seen that the concatenation can improve the performance in AD vs. NC and pMCI vs. sMCI, but with slight improvements among which the highest one is only 0.88% concerning ACC in the task of classifying pMCI from sMCI. In spite of the slight increases compared to each level of features, concatenating 3 levels of features is still not as effective as the optimal sub-type feature in pMCI vs. sMCI. Therefore, the strategy of concatenating features is not an effective method to improve the classification performance for the two tasks.

Table 4.7: Performance of different levels of feature for AD vs. NC(%)

Method	ACC	SEN	SPE	AUC
1st-Level	87.08	86.76	87.74	94.73
2nd-Level	**88.74**	**88.49**	**89.37**	**95.25**
3rd-Level	83.50	83.28	83.52	90.18
1st & 2nd & 3rd	89.39	88.76	90.13	95.41

Table 4.8: Performance of different levels of feature for pMCI vs. sMCI(%)

Method	ACC	SEN	SPE	AUC	bACC
1st-Level	67.65	61.08	71.67	72.00	66.38
2nd-Level	**70.44**	55.82	**79.19**	**74.29**	67.51
3rd-Level	68.75	**63.79**	71.83	73.81	**67.81**
1st & 2nd & 3rd	71.32	57.69	79.39	75.12	68.54

4.3.5 Effectiveness of the similarity-driven ranking method

The similarity-driven ranking method can not only reduce the 2nd-Level feature's dimension, but also improve the classifier's diversity. Here, Kappa index [130] is applied to measure the diversity and a small value indicates a high diversity, which is computed through:

$$Ka(i,j) = \frac{p_1 - p_2}{1 - p_2} \tag{4.20}$$

where p_1 denotes the observed agreement of $\mathbf{y_i}$ and $\mathbf{y_j}$, and p_2 stands for the chance agreement.

Figure 4.6 shows the effectiveness of the proposed ranking method on the diversity improvement, as can be seen, the decomposed features can achieve a higher diversity (a smaller value) than the original 2nd-Level feature for both tasks, especially for classification of pMCI. The highest diversity is attributed to connectivity $\mathbf{w}^l$, and its Kappa index with mean intensity, standard deviation, clustering coefficient and strength is 0.2367, 0.1887, 0.2981 and 0.2544, respectively. The higher diversity benefited from the similarity-driven ranking method can ensure the ensemble classifier has good performance.

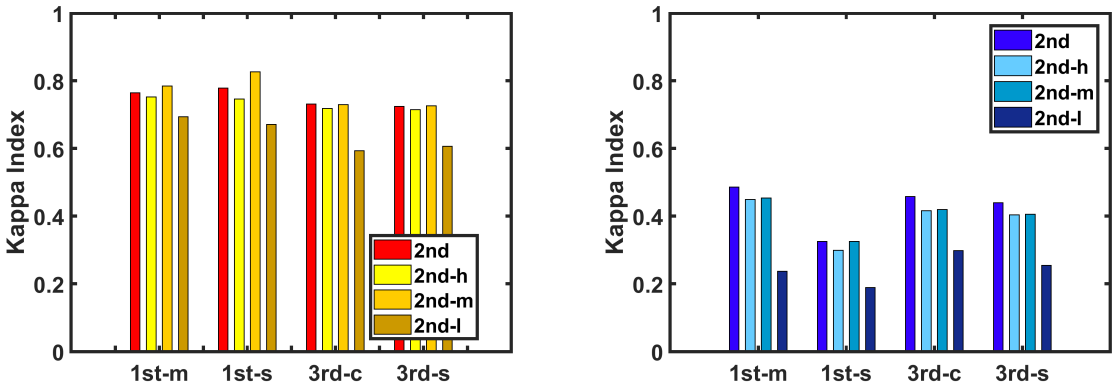
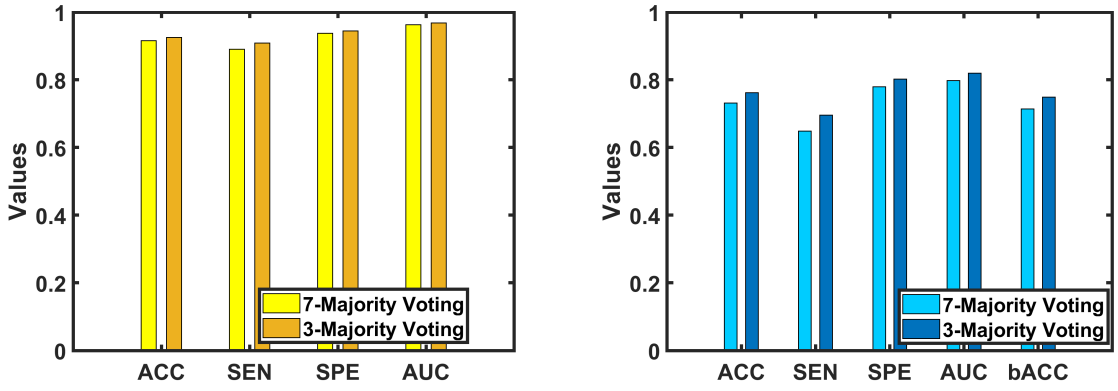


Figure 4.6: Performance evaluation of the similarity-driven ranking method. (a) AD vs. NC. (b) pMCI vs. sMCI.

4.3.6 Ensemble classification evaluation

The increase of the number of classifiers and their diversities can improve the performance of the ensemble classifier in theory. Obviously, the maximum number of classifiers (7 classifiers) is fixed in this chapter. If the sub-classifiers do not perform well and all of them are used to do the final decision through majority voting, there will be a high probability that misclassified results accounted for the majority. In order to avoid this situation and enhance the ensemble effect, a strategy of selecting models with high diversity is proposed. In this experiment, we compare majority voting using outputs from all the 7 SVMs (noted as 7-Majority Voting) with the proposed method which applies 3 selected SVMs' decisions (noted as 3-Majority Voting), and the results are shown in Figure 4.7. It can be seen that the proposed method outperforms the 7-Majority Voting, specifically, it improves by 1.02%, 1.83%, 0.56%, and 0.54% in respect of ACC, SEN, SPE and AUC in AD diagnosis and for pMCI vs. sMCI, the proposed method increases by 3.06% (ACC),

4.73% (SEN), 2.32% (SPE), 2.19% (AUC) and 3.53% (bACC). Clearly, the proposed method shows an effective improvement for classifying pMCI from sMCI, especially for SEN, which is increased by 4.73%. It implies that the ensemble classifier can improve the diagnose rate among true diseased subjects. The reason why the ensemble classifier is effective is that a single type of feature in the classification of pMCI does not perform well. The probability that misclassified results dominate the majority voting will be high, if considering all the 7 classifiers' outputs. And another reason is that the improvement of performance in classifying pMCI from sMCI benefits from the increase of diversity brought by the decomposition of 2nd-Level feature.



(a) (b)

Figure 4.7: Performance evaluation of the ensemble classification. (a) AD vs. NC. (b) pMCI vs. sMCI.

4.4 Conclusion

AD and MCI diagnoses under FDG-PET single modality are challenging. In this chapter, a novel ensemble method which uses multi-level features is proposed to address the problem. First, 3 levels of features that represent properties of ROIs and their connectivities are extracted gradually. Then a proposed similarity-driven ranking method is applied to decompose the 2nd-Level feature to 3 different sets of features, which reduces the feature dimension to a great extent and increases the classifier's diversity. Next, different models are trained by using different types of features. In order to enhance the ensemble effect, a pair of models with high diversity are selected through the proposed mMsE method and another model with high accuracy is chosen by nested cross validation. The final decision is made through the majority voting of the 3 selected models' outputs. According to experiments on

ADNI dataset, the proposed method can improve the performance of AD diagnoses and especially classifying pMCI from sMCI compared to the commonly used ROI features (mean intensity), from 71.46% to 76.17% regarding accuracy.

Chapter 5

Multiscale Spatial Gradient Features for Characterizing FDG-PET Images

5.1 Introduction

In Chapter 4, multilevel feature representation has been proposed to characterize FDG-PET images in which features are either region's properties or connectivities among regions, and such features are still ROI-wise. In fact, ROI-wise feature is in the leading place in characterizing neuroimaging data, not only for FDG-PET but also for MRI because of its relative effectiveness and less computing consumption. In this chapter, we attempt to represent FDG-PET images from the view of spatial gradients. The motivation is that the differences of glucose metabolisms between AD and NC subjects result in intensity differences in images, and also cause the differences of gradients. Therefore, it is reasonable to use spatial gradients as features in classifying AD from NC. In the remaining parts of this chapter, the proposed method is introduced in details, and then experimental results are presented and analyzed, thereby proving its effectiveness in AD diagnosis and identifying pMCI from sMCI. At last, a conclusion is given.

5.2 Method

The spatial gradient is quantified by a 2D histogram of orientation, which is similar to Histogram of Oriented Gradient (HOG) [36] that has been successfully

applied for object detection in 2D images. First, the spatial gradient of FDG-PET image is computed and then 90 ROIs are extracted from the gradient image through AAL atlas, in which the cerebellum is not considered. Next, some distinctive ROIs are selected through a proposed ROI ranking method which considers multiple Small Scale HOG (SSH) descriptors of each region. Finally, an ensemble classifier is trained under the selected ROIs by using SSH and Large Scale HOG (LSH) features. Figure 5.1 illustrates the flowchart of the proposed diagnosis method.

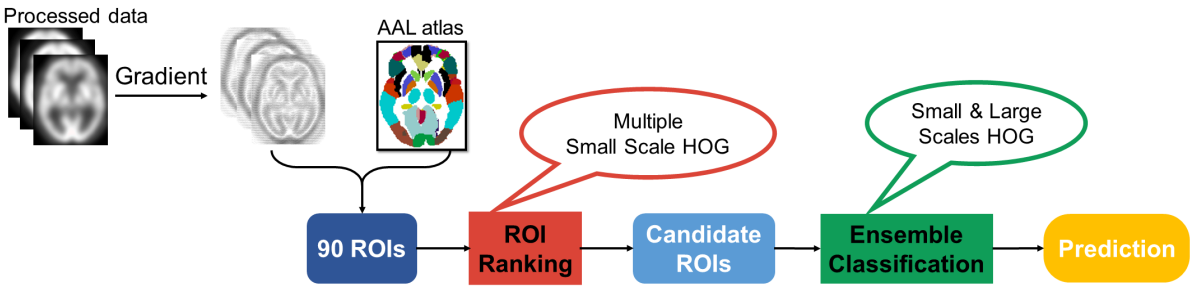


Figure 5.1: The flowchart of the proposed method.

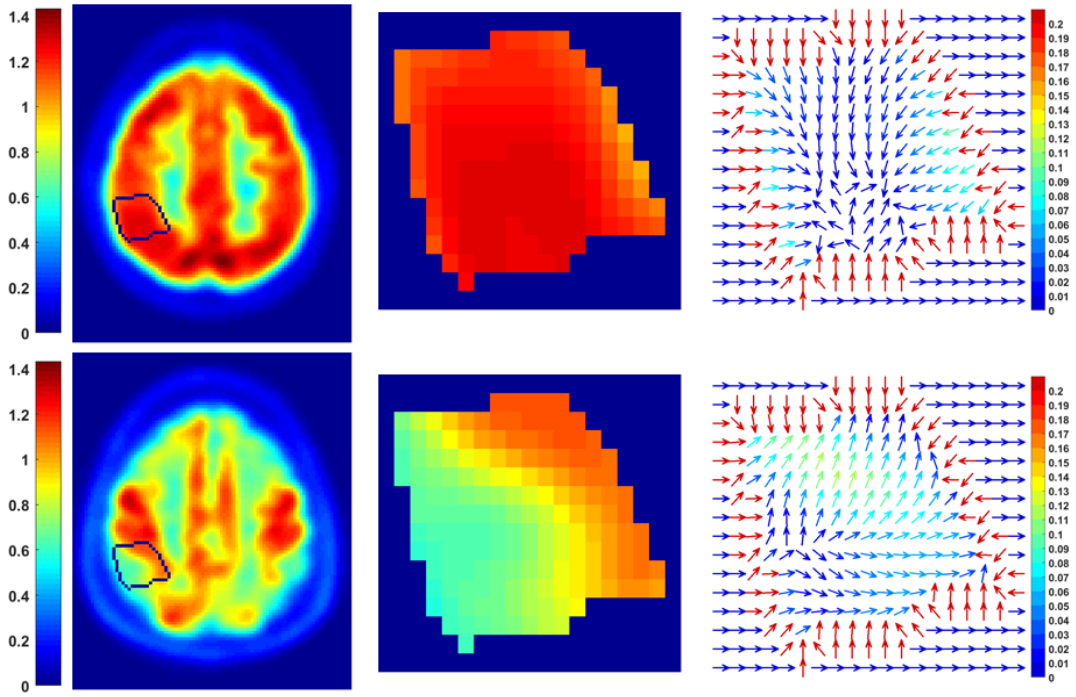
5.2.1 Histogram of Oriented Gradient for 2D images

Histogram of Oriented Gradient (HOG) is a descriptor which was proposed for human detection. The fundamental idea is that the local object appearance and shape within an image can be represented by the distribution of intensity gradients or edge directions. Generally, the image is divided into small connected regions, and for the pixels within each region, a histogram of gradient directions is computed. The descriptor is the concatenation of these histograms. HOG feature is an effective hand-crafted descriptor for object detection since it can capture the edge or gradient structure which is discriminative for local shape [36].

5.2.2 Histogram of Oriented Gradient for FDG-PET images

Figure 5.2 shows the difference between an NC (top row) and an AD (bottom) subjects. Figure 5.2(a) displays one of the slices and the circled area belongs to region Parietal_Inf_R in AAL template. The enlarged areas are shown in Figure 5.2(b) (different colors indicate different intensities). It can be clearly seen that the intensities are different between NC and AD PET scans. In addition, the gradients are also different, as shown in Figure 5.2(c) (different colors indicate different gradient magnitudes), and the gradient change of the AD scan (bottom) is more obvious

than that of NC (top). This observation drove us to investigate the effectiveness of spatial gradients in diagnosing AD. In the following, we exploit a 2D histogram of orientation to quantify spatial gradients in order to characterize FDG-PET images.



(a) (b) (c)

Figure 5.2: An instance to differentiate between NC and AD via intensity and gradient, where the top row is for an NC subject and the bottom row is for a subject with AD. (a) A slice of a subject. (b) Part of region Parietal_Inf_R. (c) The corresponding gradient.

Spatial Gradient Computation For a 2D image, the gradient is computed from the horizontal and vertical directions, and the corresponding orientation is determined by one angle. Similarly, the spatial gradient of a 3D image is calculated in the x, y and z directions and the orientation is decided by two angles.

For a voxel with an intensity $f(x, y, z)$ at the position (x, y, z), its numerical gradient can be computed as:

$$\begin{aligned} g_x &= 0.5 \times (f(x+1, y, z) - f(x-1, y, z)) \\ g_y &= 0.5 \times (f(x, y+1, z) - f(x, y-1, z)) \\ g_z &= 0.5 \times (f(x, y, z+1) - f(x, y, z-1)) \end{aligned} \tag{5.1}$$

where g_x, g_y and g_z are gradients in the x, y and z directions, respectively. The

magnitude $\|\mathbf{g}\|$ is obtained through:

$$\|\mathbf{g}\| = \sqrt{{g_x}^2 + {g_y}^2 + {g_z}^2} \tag{5.2}$$

and the orientation is represented by the polar angle, ϕ, and the azimuth angle, θ, as shown in Figure 5.3

$$\begin{aligned} \phi &= \arctan\left(\frac{g_y}{g_x}\right) \\ \theta &= \arccos\left(\frac{g_z}{\|\mathbf{g}\|}\right) \end{aligned} \tag{5.3}$$

where ϕ is in the range $[-180°, 180°]$ and θ is in the range $[0°, 180°]$.

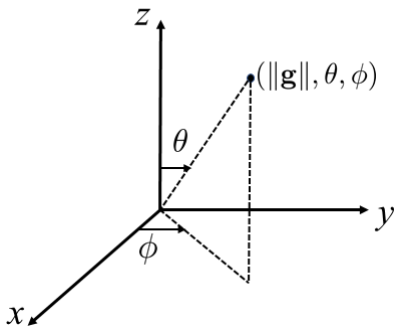


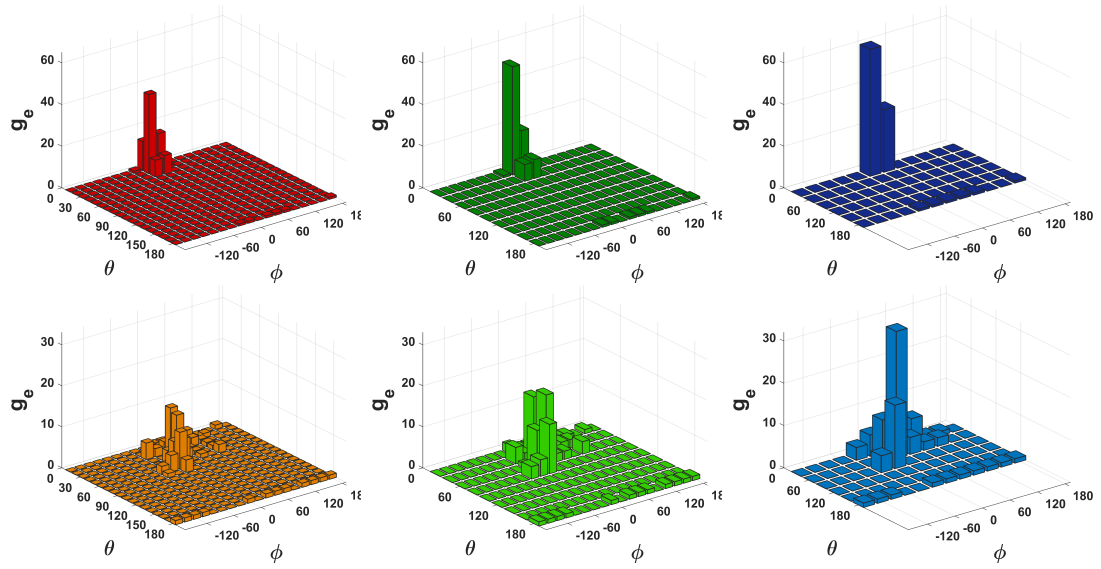
Figure 5.3: An instance of the polar angle ϕ and azimuth angle θ.

Spatial Gradient Quantification In this step, a 2D histogram is constructed based on gradient orientations (ϕ and θ), and the magnitude is used to count the occurrence of a certain orientation. Specifically, ϕ and θ are viewed as two properties of the 2D histogram, and then are evenly divided into several intervals or bins, respectively. Last, if the gradient orientation is within a certain interval, the *value* for that interval is accumulated. According to [36], the *value* is computed via a function of the gradient magnitude. Considering the magnitude of each voxel is small, the exponential function of magnitude, g_e, is applied as the counting *value* in each interval,

$$g_e = \exp(\|\mathbf{g}\|) \tag{5.4}$$

Consequently, the FDG-PET image can be represented by a 2D histogram, and meanwhile, the representation will vary with the number of bins in the histogram. Figure 5.4 shows histograms with different numbers of bins for the same FDG-PET image, where the top row is for an NC subject and the bottom row is a subject with

AD. It can be seen that for a subject, either NC or AD, a change in the number of bins can result in different representations (each row). Moreover, the difference between NC and AD is obvious under the same number of bins (each column), which also implies that it is reasonable to use 2D HOG as the feature to diagnose AD.



(a) (b) (c)

Figure 5.4: Histogram of Oriented Gradient (HOG) for an NC subject (top row) and an AD subject (bottom row) without segmentation. (a) 18×18 ($\phi \times \theta$) bins. (b) 18×9 bins. (c) 12×6 bins.

Compared with Global HOG which is computed on the whole subject without segmentation, calculating locally is also worth considering. To achieve this goal, each subject is divided into 90 regions by using AAL atlas, and a histogram can be constructed for each region. It should be noted that the 2D histogram is computed in irregular regions which contain different numbers of voxels. Since the number of bins in a histogram is adjustable, multiple scales of 2D HOG features are extracted from each ROI to make the features informative, including 18×18 ($\phi \times \theta$), 12×18, 18×9, 12×9 and 12×6 bins, which are denoted Small Scale HOG (SSH) because of the small interval used in gridding angles for constructing a histogram. Correspondingly, some Large Scale HOG (LSH) descriptors, in which a large interval is used to grid angles, are also extracted from each ROI, including 5×5 ($\phi \times \theta$), 1×1 bins. The features from different ROIs exhibit different ranges, so normalization is essential in order to achieve good performance. In this work, L_2 norm is used to do the normalization.

$$\mathbf{h}' = \frac{\mathbf{h}}{\|\mathbf{h}\|} \tag{5.5}$$

where $\mathbf{h}$ is 2D HOG vector for each scale of any ROI in a subject, $\mathbf{h}'$ stands for 2D HOG descriptor after normalization.

5.2.3 ROI ranking

The probability of AD occurring in each ROI is not consistent, which means that some ROIs are more likely to be affected by AD, while others are not. In practice, doctors pay more attention to key areas as well. Therefore, it is necessary to select typical ROIs which are more susceptible to AD. A region ranking method is developed to achieve the goal of region selection. Specifically, 2D HOG features are applied to characterize each ROI and then fed into a linear SVM to compare each ROI's classification accuracy (or balanced accuracy for unbalanced situation), thereby ranking ROIs. In order to obtain a robust and reliable result, multiple SSH features are considered, and the average accuracy is used for ranking ROIs. The reason LSH is not taken into account is that the feature dimension of LSH is relatively low, for example, the dimension of 1×1 LSH is only 1. Figure 5.5 shows the framework of the ranking method. A region with a higher classification accuracy implies its stronger ability to recognize AD. At last, some ROIs can be selected through the ranking order. The ranking results are analyzed in Section 5.3.3.

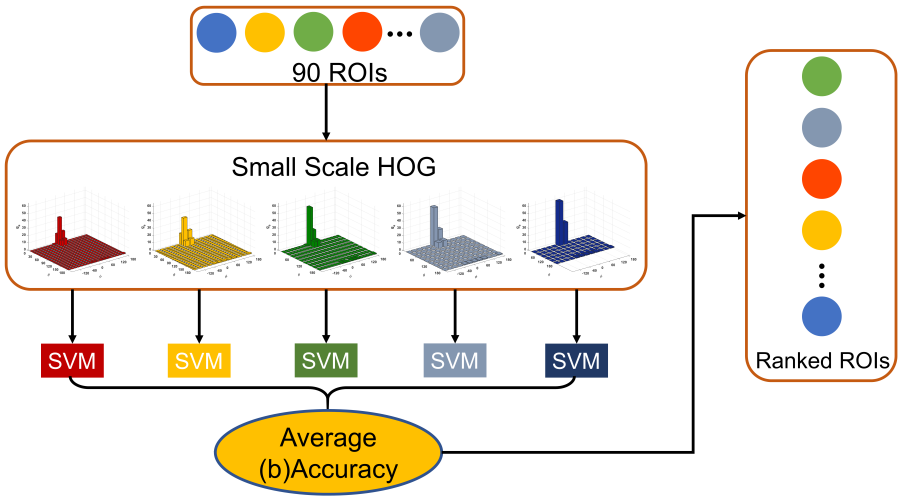


Figure 5.5: The framework of region ranking method.

5.2.4 Ensemble classification

After ranking ROIs, the top N ROIs with higher performance are selected as candidate regions for the classification. SVM is a popular and effective classifier in

AD diagnosis. In this study, SVM is applied in an ensemble classification framework which considers both SSH and LSH features of selected ROIs. The motivation of designing an ensemble framework is inspired by the idea that weak classifiers can be combined into a strong classifier. Although SVM is usually treated as a strong classifier, we make it become a relatively weak model through controlling input features (ROI by ROI, scale by scale) and setting the parameter C which is introduced in Section 5.3.1. In addition, concatenating all the scales of 2D HOG and all the regions is not an efficient way to train a model due to a large amount of inputs. Therefore, the selected regions guided by the ROI ranking method are utilized to train a set of weak classifiers, and then multiple classifiers are integrated to make a prediction. The framework of the ensemble classifier is briefly presented in Figure 5.6. Specifically, five types of SSH features (18×18 ($\phi \times \theta$), 12×18, 18×9, 12×9 and 12×6 bins) are extracted from each selected ROI and then are used to train five classifiers, respectively. The average score of five classifiers is considered as the corresponding ROI's output, $\mathbf{S}$, which is expressed as:

$$\mathbf{S} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{s}_t \tag{5.6}$$

where $\mathbf{s}_t$ is the output score of SVM with t-th scale of SSH descriptor and T is the number of scales of SSH, here $T = 5$. For LSH, all the features of candidate ROIs are concatenated to feed into a classifier because of the low feature dimension. The final decision, $\mathbf{Y}$, is made through an addition strategy of two parts' outputs, small and large scales,

$$\mathbf{Y} = \text{sgn}\left(\sum_{i=1}^{N} \mathbf{S}_i + \sum_{j=1}^{M} \mathbf{L}_j \right) \tag{5.7}$$

where $\text{sgn}(\cdot)$ is a sign function, N is the number of candidate ROIs which is decided via experiments, M is the number of scales of LSH, $M = 2$, and $\mathbf{L}$ is the score of LSH-based classifier. Therefore, the proposed ensemble classifier not only considers the performance of each individual ROI, but also considers the performance of cascade regions.

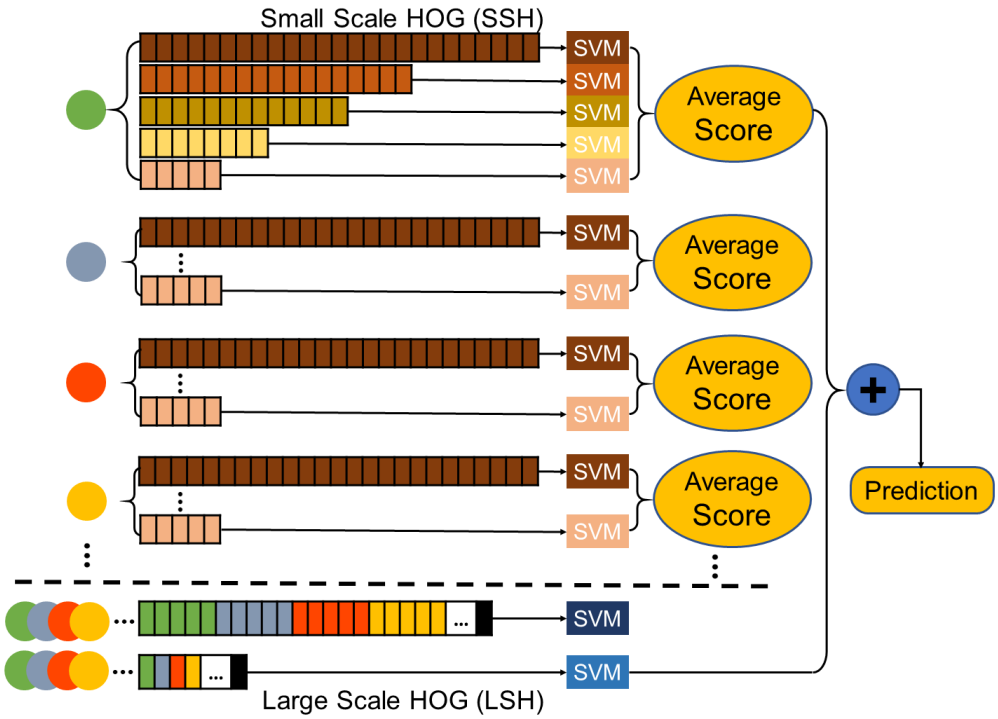


Figure 5.6: The framework of the ensemble classification, circles with different colors indicate different regions.

5.3 Experiments and results

5.3.1 Setup

Similar to experiments in Chapter 4, experiments in this chapter are conducted on two classification tasks, AD vs. NC and pMCI vs. sMCI as well. Four metrics, classification accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under curve (AUC) are applied to evaluate the corresponding performance and an extra measurement, balanced ACC (bACC) for the unbalanced situation. For all the metrics, a higher value indicates better performance. Due to the limited number of subjects, we use a 10-fold cross-validation technique to assess the performance and repeat 10 times to reduce the possible bias. The margin parameter C of all the SVMs used in the ensemble classifier is set to 0.5 in order to construct relatively weak classifiers.

5.3.2 Evaluation on spatial gradient features for FDG-PET images

The spatial gradient feature is compared with commonly used voxel-wise and ROI-wise features. The comparison results of different representations are presented in Table 5.1 and Table 5.2 for AD vs. NC and pMCI vs. sMCI, respectively. The terms 'Voxel' and 'ROI' stand for classification results which are obtained by using voxel intensity and region's mean intensity, respectively. 'Global HOG' means 2D HOG descriptor is computed on the subject without parcellation. The results of Global HOG and SSH shown in the following two tables are achieved based on a histogram with 18×9 bins, while for LSH, its results are computed on a histogram with 5×5 bins.

Table 5.1: Comparison of common features and spatial gradient features for AD vs. NC(%)

Feature	Dimension	ACC	SEN	SPE	AUC
Voxel	**160990**	92.83	91.90	93.71	97.17
ROI	90	88.56	87.08	90.00	94.91
Global HOG	162	62.80	60.40	65.83	67.94
LSH	2250	93.63	92.01	95.50	97.97
SSH	14580	**94.53**	**92.43**	**96.60**	**98.22**

Table 5.2: Comparison of common features and spatial gradient features for pMCI vs. sMCI(%)

Feature	Dimension	ACC	SEN	SPE	AUC	bACC
Voxel	**160990**	70.85	58.90	77.88	75.34	68.39
ROI	90	72.37	58.90	80.28	77.36	69.59
Global HOG	162	57.47	56.32	58.54	58.52	57.43
LSH	2250	72.73	**59.91**	80.33	76.80	70.12
SSH	14580	**74.92**	57.63	**85.08**	**79.97**	**71.36**

It can be seen from Table 5.1 that voxel intensity is a kind of effective feature in classifying AD from NC, which achieves an accuracy of 92.83%. But the feature dimension is too large, which is a drawback for training a model. Even though the dimension of ROI-wise feature is small, its performance is not significant and it is not as effective as voxel-wise feature, with an accuracy rate of 88.56%. Global HOG,

with 62.80% accuracy, cannot achieve a noteworthy result. It can be explained that: 1) Global HOG is computed on the whole subject without considering local details; 2) the dimension, 162 (18×9), seems to be a satisfactory one, but the effective information is much less than that because of a lot of zero values, like Figure 5.4(b), which is less informative. LSH and SSH descriptors can guarantee the performance and meanwhile tackle the problem of large dimension. Even though 14580 ($18 \times 9 \times 90$) is not an absolutely desirable dimension, it is still acceptable compared to the dimension of the voxel-wise feature. In addition, owing to the adjustable number of bins in a histogram, a smaller dimension can be obtained if setting an appropriate number of bins.

As for classifying pMCI from sMCI, which is reported in Table 5.2, voxel-wise feature is inferior to ROI-wise feature, which is different to their performance in AD classification. The reason could be that the dimension of voxel-wise feature is far greater than that of ROI-wise feature, which can cause the problem of feature redundancy. Such a problem would harm the classifier training, especially for the case that the classifier could not work well. As can be seen from Table 5.2, all the features (voxel-wise, ROI-wise, Global HOG, LSH and SSH) cannot perform as well as in AD diagnosis, thus prediction of pMCI is a challenging task and it is easily influenced by redundant features. Nevertheless, SSH descriptor still has dominant performance with a balanced accuracy of 71.36%, which is 1.77% higher than the ROI-wise feature, and LSH has comparative performance with the ROI-wise features. The other three metrics, including ACC, SPE and AUC, also indicate SSH is superior to the other features. In summary, characterizing FDG-PET images by 2D HOG locally is effective and feasible. The reason could be that the spatial gradient is calculated at voxel level and the 2D histogram is computed at ROI level, which makes the 2D HOG descriptor become a *bridge* to link voxel-wise feature and ROI-wise feature.

5.3.3 Evaluation on ROI ranking method

Since different regions have different abilities to diagnose AD, a simple ROI ranking method using multiple SSH features and SVM is proposed. Specifically, for each region, five scales of SSH features are extracted and then fed into five SVMs, respectively. The average accuracy of classifiers is considered as the ranking metric, and the higher the accuracy is, the stronger the ability of the region to distinguish AD from NC. Considering pMCI vs. sMCI is an unbalanced case, balanced accuracy

is exploited as the metric instead accuracy. Multiple SSH features are applied to ensure the reliability and robustness of the results. Figure 5.7(a) shows the maximum difference of accuracy (or balanced accuracy) in five scales of SSH, Δacc, for each ROI. As can be seen, the difference is obvious, especially for identifying pMCI from sMCI, and the highest Δacc is 10.05%. The maximum difference of area under curve, Δauc, is illustrated in Figure 5.7(b). It can be seen that changes among different scales are significant as well, particularly for Δauc in identifying pMCI from sMCI, which proves that it is rational to use multiple SSH features to rank ROIs. Moreover, the change of maximum difference in pMCI prediction is larger than that of AD classification, which is because that the latter task is easier than the former one. For AD classification, different scales of effective features usually achieve high-level accuracies or other metrics but with a little difference. While for pMCI prediction, it is challenging, so different scales of features may not be that effective and may achieve unstable performance, which could cause a larger difference within different scales of features.

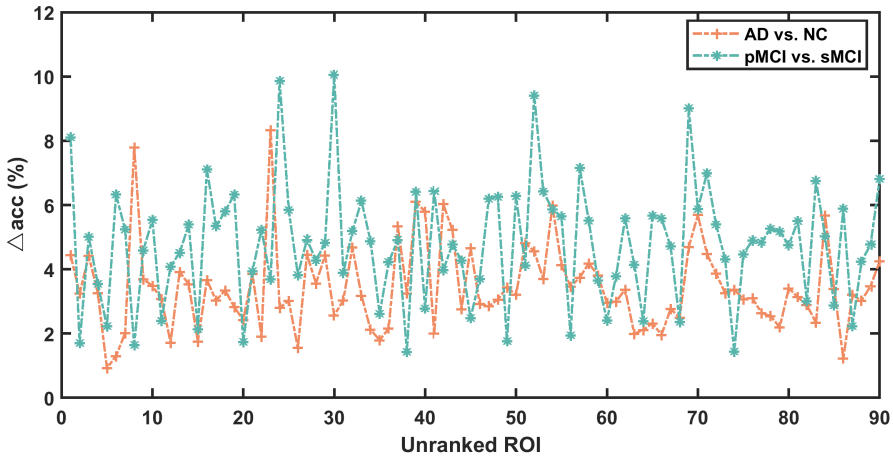
Table 5.3 presents the top 20 regions ranked by the proposed method for two tasks, AD vs. NC and pMCI vs. sMCI, which can be regarded as potential FDG-PET indicators for the subsequent classification tasks. Compared to regions selected by multilevel feature representation method which are listed in Table 4.1, there is a certain amount of ROIs that both methods have suggested, such as ROIs Hippocampus_R/L, Fusiform_R/L, Temporal_Inf_L, Precuneus_R, *etc.* for AD diagnosis, and ROIs Precuneus_R/L, Temporal_Mid_R, Parietal_Sup_R, etc. for pMCI vs. sMCI.

5.3.4 Evaluation on ensemble classification

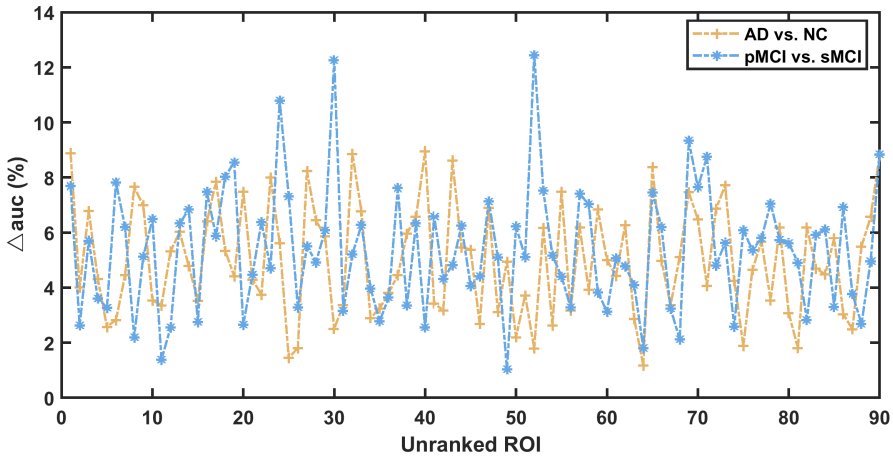
An ensemble classification framework is designed by considering SSH and LSH descriptors together since these multiple scales of 2D HOG contain both specific (SSH) and general information (LSH). Figure 5.8 shows (balanced) accuracies under different numbers of regions for AD vs. NC and pMCI vs. sMCI. It should be noted that the x axis denotes ROIs which have been ranked from most to least relevant to AD/pMCI according to the proposed ROI ranking method. As can be seen from Figure 5.8(a), SSH achieves its highest ACC, 93.25%, with fewer regions (16 ROIs) than LSH whose best performance (ACC: 92.85%) is obtained by using 82 ROIs. Thus SSH descriptor performs slightly better than LSH in AD diagnosis. In addition, the ensemble classification through integrating SSH and LSH improves

Table 5.3: Top 20 ROIs for AD vs. NC and pMCI vs. sMCI

Rank	AD vs. NC	pMCI vs. sMCI
1	Cingulum_Mid_R	Cingulum_Mid_L
2	Hippocampus_R	Parietal_Inf_R
3	Cingulum_Mid_L	Parietal_Inf_L
4	Hippocampus_L	Occipital_Sup_R
5	Fusiform_R	SupraMarginal_R
6	Precuneus_R	Precuneus_R
7	Cuneus_R	Precuneus_L
8	Temporal_Inf_L	Cuneus_R
9	Precuneus_L	Temporal_Mid_R
10	Parietal_Inf_R	Cuneus_L
11	Cuneus_L	Temporal_Sup_L
12	Paracentral_Lobule_R	Fusiform_R
13	Paracentral_Lobule_L	Hippocampus_R
14	Temporal_Mid_L	Occipital_Mid_R
15	Fusiform_L	Parietal_Sup_R
16	Temporal_Sup_L	Calcarine_L
17	Parietal_Sup_R	Paracentral_Lobule_R
18	Occipital_Sup_R	Occipital_Mid_L
19	Angular_L	Cingulum_Mid_R
20	Occipital_Mid_R	Temporal_Mid_L



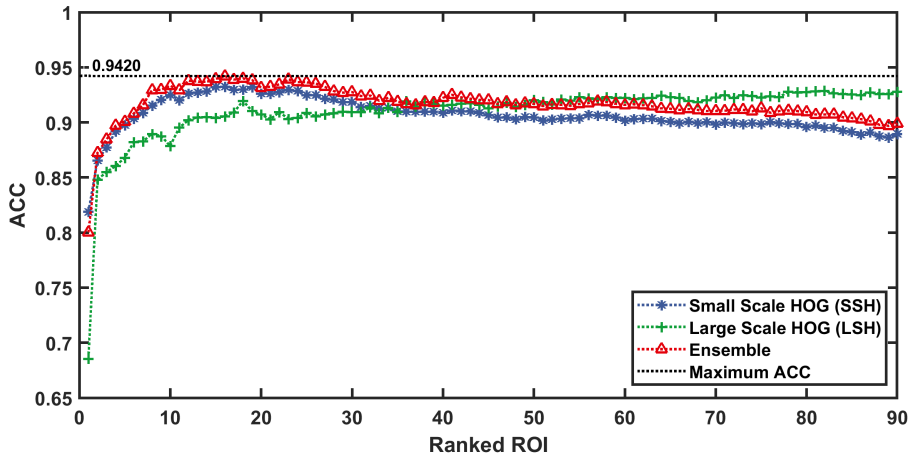
(a)



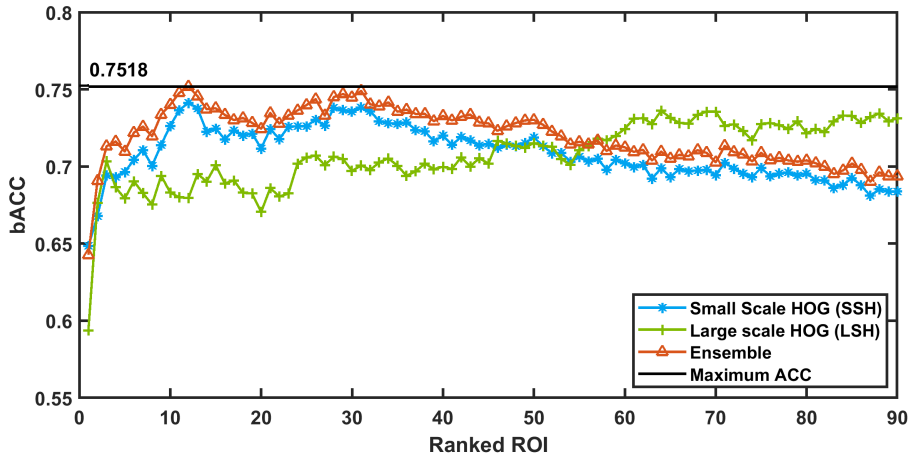
(b)

Figure 5.7: Maximum difference of accuracy (or balanced accuracy) (a), and area under ROC (b), in five scales of SSH for each ROI

the accuracy to 94.20% under top 16 ROIs, which achieves a slight increase for AD diagnosis (0.95%). In the case of pMCI vs. sMCI, as illustrated in Figure 5.8(b), the performance of SSH is still slightly better than LSH, specifically, SSH reaches the best balanced accuracy, 74.14% with top 12 regions, while LSH achieves its best bACC, 73.62% by using top 64 regions. Furthermore, the ensemble classifier raises the accuracy to 75.18% with 12 ROIs, and can significantly improve performance under different numbers of ROIs. In summary, for the two tasks (AD vs. NC and pMCI vs. sMCI), LSH descriptor does not perform as well as SSH feature when employing less regions. It is because LSH contains more general features, which makes it less informative than SSH. However, more regions can strengthen the performance of LSH. Besides, the ensemble classifier is better than the SSH-based classifier, which proves that the integration strategy is effective.



(a)



(b)

Figure 5.8: Performance under different numbers of ROIs. (a) AD vs. NC. (b) pMCI vs. sMCI.

5.4 Conclusion

This chapter extends the feature in object detection, HOG, to FDG-PET brain images to aid AD diagnosis, which is effective as well. Compared to the classic representations (voxel-wise and ROI-wise), 2D HOG descriptor is more informative than the ROI-wise feature and more sparse than the voxel-wise feature and meanwhile, can ensure effectiveness. Besides, a ROI ranking method is proposed by applying multiple SSH descriptors and according to which, a set of candidate ROIs are selected to assist the diagnosis. Furthermore, an ensemble classification framework is designed over the selected regions through using SSH and LSH features. The ensemble classifier is effective and outperforms other methods according to the evaluation on the ADNI dataset.

Chapter 6

Multiview Convolutional Neural Network for AD diagnosis and MCI Conversion Prediction

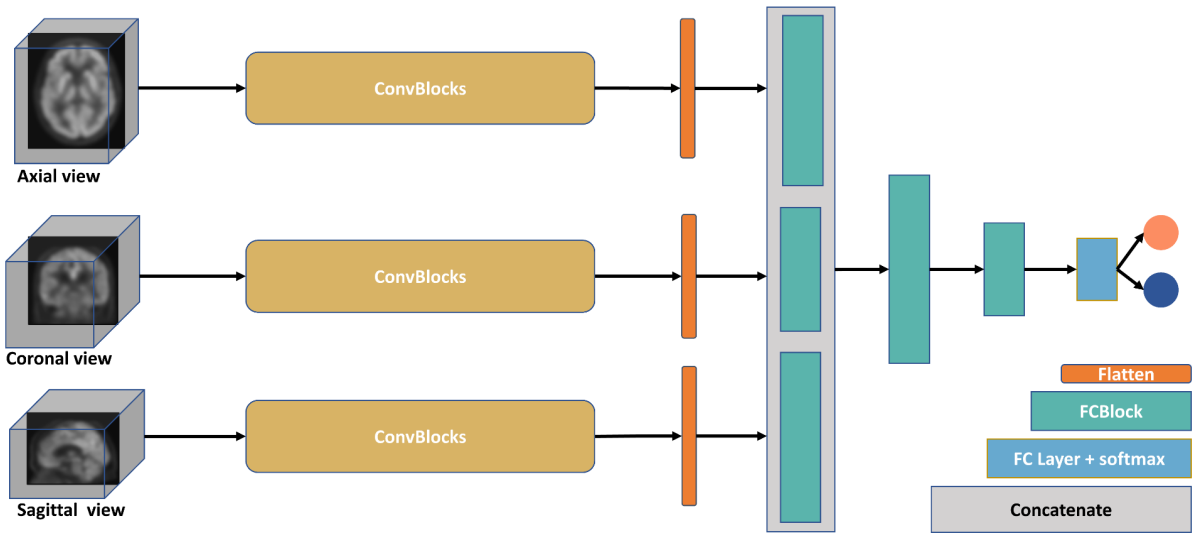
6.1 Introduction

In this chapter, deep learning techniques, mainly the Convolutional Neural Network (CNN), are applied to tackle problems of AD diagnosis and MCI conversion prediction under the FDG-PET modality. Generally, a 3D CNN is used in neuroimaging data since the image has three dimensions, but too many parameters are involved in such methods. A 2D CNN can also be applied to address the diagnosis problem in which the CNN works slice by slice, but the spatial relations among voxels are not taken into account. Therefore, in order to reduce parameters involved and meanwhile consider the spatial relations, a novel method which uses a multiview CNN architecture is proposed. In the remaining of the chapter, the proposed method, mainly the network structure and its corresponding implementations, is firstly described. Then different evaluations are taken to assess the performance of the proposed method.

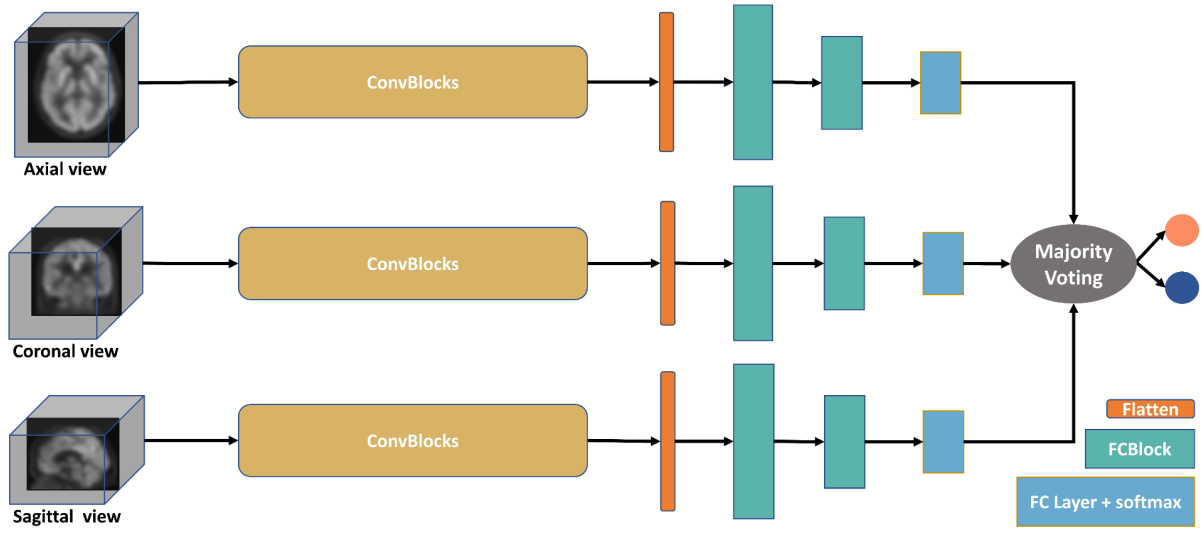
6.2 Method

Two types of multiview CNN architectures are proposed, as illustrated in Figure 6.1. It can be seen that axial, coronal and sagittal views are performed CNN at first, respectively. Three views are then combined to yield results jointly. The main

difference between the two types of architectures lie in the combination manner. One is to concatenate multiple views at the first fully connected (FC) layer of each branch and then pass another three FC layers prior to make a decision, denoted mvCNNiF, as shown in Figure 6.1(a). The other one is to integrate results from multiple views after the last FC layer of each branch through a majority voting fashion, denoted mvCNNaF, as shown in Figure 6.1(b). Accordingly, three views involved in mvCNNiF are trained simultaneously, while for mvCNNaF, models along axial, coronal and sagittal views are trained separately and then make a decision jointly.



(a)



(b)

Figure 6.1: The proposed multiview CNN architectures. (a) mvCNNiF. (b) mvCNNaF.

6.2.1 Multiview CNN architecture

Before being delivered to a convolutional block, each view is projected to a 2D image by performing an extra convolution operation, which is displayed in Figure 6.2. As can be seen, the convolutional kernel is like a cuboïd whose size is $1 \times 1 \times 91$ for axial and sagittal views and $1 \times 1 \times 109$ for coronal view instead of a commonly used size, like $3 \times 3 \times 3$, $5 \times 5 \times 5$. This operation can map the spatial information along each dimension onto its corresponding plane, thereby achieving the goal of dimension reduction. This special convolutional layer is defined as the mapping layer. The motivation behind using 1×1 for the first two dimensions of a convolutional kernel is to ensure the mapping operation is only performed on the third dimension and the output size along the first two dimensions remains unchanged and meanwhile, the size of 1×1 can also reduce parameters relatively compared to 3×3. As a result, the outputs of mapping layers are 109×91, 91×91 and 91×109 for axial, coronal and sagittal views, respectively. It is then followed by batch normalization (BN) [131] which enables to normalize the outputs of a layer by subtracting their average value and dividing by the corresponding standard deviation. This procedure can enforce a fixed distribution of activations, thereby stabilizing and accelerating the training of deep neural networks. Then an activation, rectified linear unit (ReLu), is applied to increase the nonlinearity between layers considering its good performance, like simplifying the computation, avoiding the gradient vanishing problem.

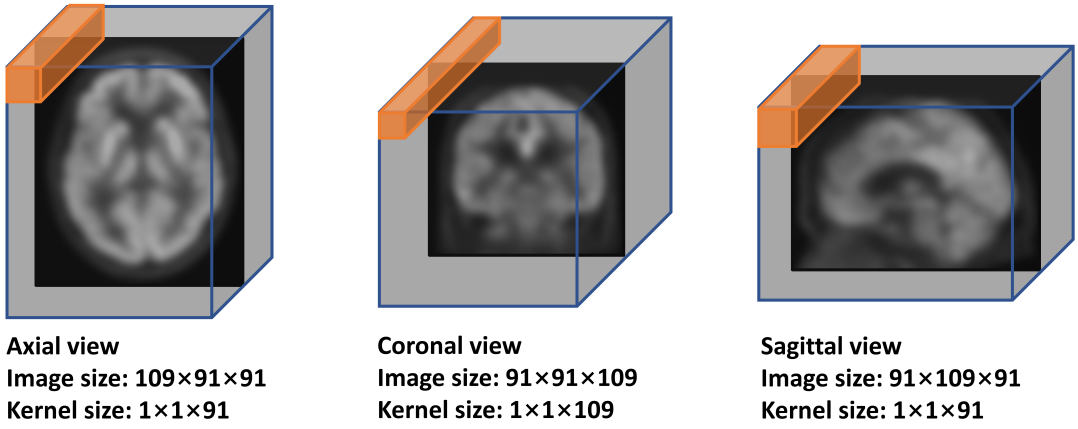
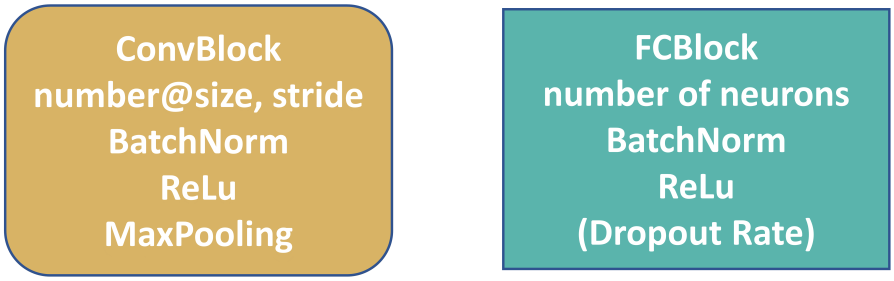


Figure 6.2: Cuboïd kernel for each view.

The mapped image from each view is then passed through several convolutional blocks, denoted ConvBlock. Each block follows a similar pattern, as displayed in Figure 6.3(a). Specifically, the ConvBlock comprises a convolutional layer which is specified by the number of kernels, its corresponding kernel size and sliding stride.

Then it is followed by batch normalization and ReLu activation. In order to reduce the resulting dimension, max pooling is then exploited, where only the maximum value within a window is retained. The corresponding window size is fixed to 2×2 and the stride is set to 2 for all the max pooling layers in this thesis. Padding is set to 0 in all the convolutional blocks. After ConvBlovks, the FCBlock, as shown in Figure 6.3(b), is performed. It consists of an FC layer with a specified number of neurons, which is then followed by BN and ReLu successively as well. After that, dropout [132] strategy is employed to avoid overfitting, which works by randomly dropping some neurons and their connections during training.



(a) (b)

Figure 6.3: Details of blocks. (a) ConvBlock. (b) FCBlock.

The specific network architecture is shown in Figure 6.4, as can be seen, for the mapping layer of each view (the first column), 8 cuboïd kernels are applied and each kernel can convert a 3D image to a 2D one. Consequently, eight 2D images or feature maps are derived, as shown in Figure 6.5, the mapping procedure is displayed intuitively. Then four ConvBlocks are deployed successively. For each block, its number of kernels, corresponding kernel size and the sliding stride are indicated in Figure 6.4. For instance, '16@5×5, 2' in ConvBlock1 implies that 16 kernels with a size of 5×5 are exploited and the sliding stride is set to 2 and ConvBlock1 can yield 16 feature maps. For the other 3 ConvBlocks (ConvBlock2, ConvBlock3, ConvBlock4), the number of kernels is set to 32, 64 and 128, respectively, and the kernel size, as well as the sliding stride, are fixed to 3×3 and 1, respectively. The outputs of ConBlock4 from three views are flattened into a vector and then integrated. There are two different ways to combine the three views, and different combination manners generate different architectures, mvCNNiF and mvCNNaF, as displayed in Figure 6.4(a) and (b). For the former one, mvCNNiF, the flattened vector of each view is followed by a FCBlock which is specified by the number of neurons and the dropout rate. For example, the axial and sagittal views in mvCNNiF (Figure 6.4(a)),

the FCBlock is with 512 neurons and a dropout rate of 0.5. For for the coronal view, 216 neurons are used in the FCBlock after a flattening operation. Then outputs of the FCBlocks are concatenated and fed into two consecutive FCBlocks with 512 (or 256 for the coronal view) and 64 neurons, respectively. Lastly, the decision is given by a 2-neuron-FC layer equipped with an activation of softmax. As to mvCNNaF, which combines three views after the last FC layer, each view undergoes 2 FCblocks with 512 (or 256 for the coronal view) and 64 neurons and a 2-neuron-FC layer with softmax. Three views are trained separately compared to those in mvCNNiF trained simultaneously. At last, a majority voting strategy is utilized to integrate the outputs of three views, thereby yielding a prediction. It should be noted that all the values within kernels, as well as neurons, are learned by the machine itself.

The architecture hyperparameters in mvCNNaF and output size are shown in Table 6.1 and Table 6.2 for axial and coronal views, respectively, and sagittal view shares the same architecture with axial view. The terms 'Conv' and 'MaxP' indicate a convolutional layer and a max pooling layer involved in a ConvBlock. The output size is computed by Eq. 2.37 in Chapter 2. For instance, the output size of ConvBlock1-Conv is calculated as,

$$\frac{109-5+2\times 0}{2}+1=53, \qquad \frac{91-5+2\times 0}{2}+1=44, \tag{6.1}$$

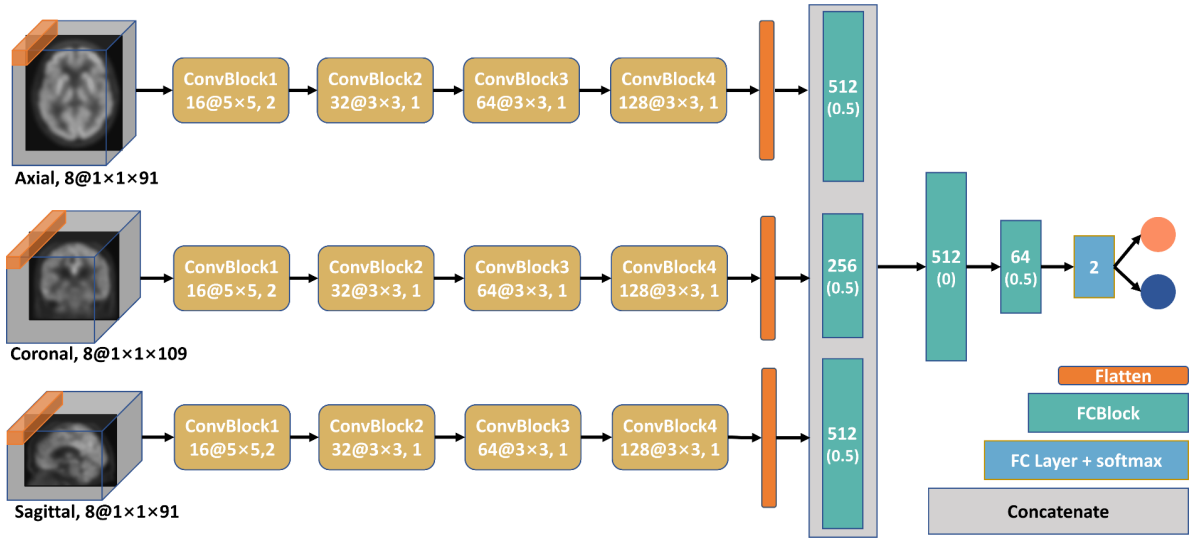
and for ConvBlock1-MaxP, its corresponding output is computed as,

$$\frac{53-2+2\times 0}{2}+1\approx 27, \qquad \frac{44-2+2\times 0}{2}+1=22. \tag{6.2}$$

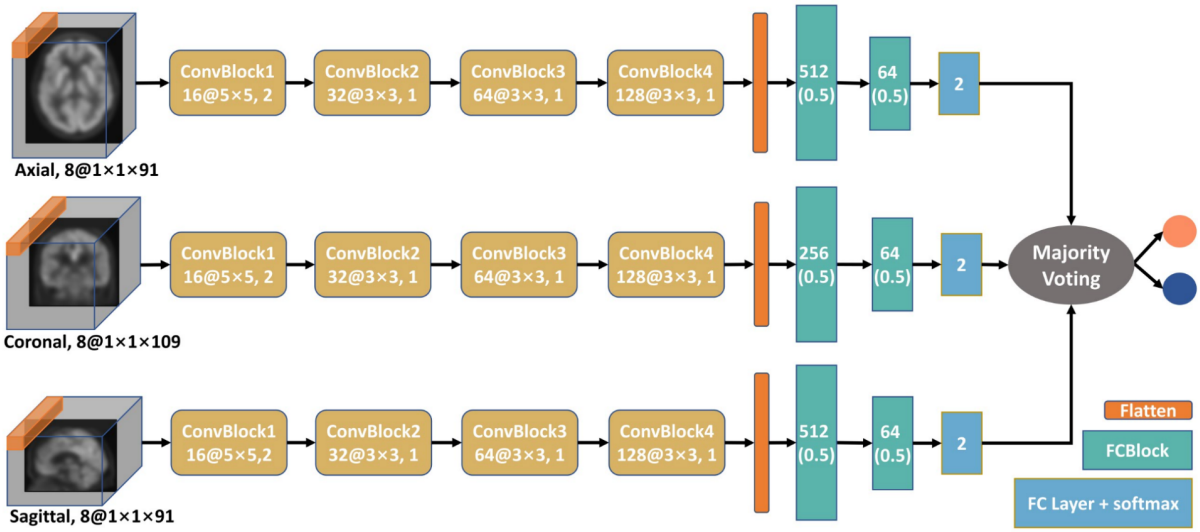
Note that hyperparameters and output size in mvCNNiF are nearly the same with those in mvCNNaF, therefore only parameters involved in mvCNNaF are given.

6.2.2 Implementations

All experiments are conducted by using python 3.6 on a Linux machine equipped with a Nvidia Quadro P5000 graphics card with 16 GB. The neural network is built with Keras deep learning library [133] using TensorFlow [134] as backend. Across all experiments, certain network settings remain unchanged. These include the dropout rate, which is set to 0.5 for all the FCBlocks except for the second FCBlock in mvCNNiF which is set to 0; and the initialization method for all the layers, which follows 'he_uniform' [135]; and batch size and number of epochs, which are fixed to 8 and 150, respectively. The objective function is minimized by a stochastic



(a)



(b)

Figure 6.4: The proposed multiview CNN architectures. (a) mvCNNiF. (b) mvCNNaF.

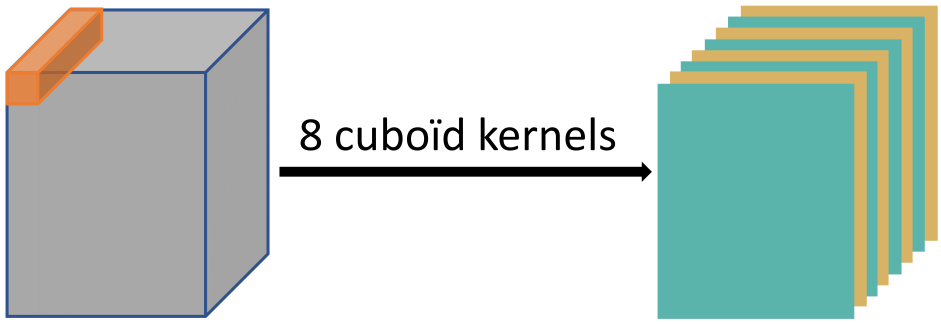


Figure 6.5: An instance of the mapping procedure.

Table 6.1: mvCNNaF architecture hyparameters and output size for **axial and sagittal** views

Layer/Block	Kernel size	Kernels/ neurons	Stride size	Dropout rate	Output size
Mapping layer	1×1×91	8	1	–	109×91
ConvBlock1-Conv	5×5	16	2	–	53×44
ConvBlock1-MaxP	2×2	–	2	–	27×22
ConvBlock2-Conv	3×3	32	1	–	25×20
ConvBlock2-MaxP	2×2	–	2	–	13×10
ConvBlock3-Conv	3×3	64	1	–	11×8
ConvBlock3-MaxP	2×2	–	2	–	6×4
ConvBlock4-Conv	3×3	128	1	–	4×2
ConvBlock4-MaxP	2×2	–	2	–	2×1
Flatten	–	–	–	–	256
FCBlock1	–	512	–	0.5	512
FCBlock2	–	64	–	0.5	64
FC3	–	2	–	–	2

gradient descent (SGD) algorithm [136] with step-wise learning rate, specifically, for AD diagnosis, 10^{-3} is set from epoch 1 to epoch 50, 10^{-4} is for epoch 51–100, and 10^{-5} is for epoch 101–150, as to pMCI vs. sMCI, the learning rate is set to 10^{-4} from epoch 1 to epoch 100 and 10^{-5} is set for the remaining epochs. The momentum coefficient is empirically set to 0.9 for both tasks. These settings are either empirically or experimentally.

6.3 Experiments and results

6.3.1 Setup

For the evaluation of the proposed multiview CNN models, experiments are still conducted on two tasks, AD vs. NC and pMCI vs. sMCI, and five metrics are utilized, including ACC, SEN, SPE, AUC and bACC which is for the unbalanced situation. The hold-out splitting strategy is employed to randomly divide the dataset

Table 6.2: mvCNNaF architecture hyparameters and output size for **coronal** view

Layer/Block	Kernel size	Kernels/ neu-rons	Stride size	Dropout rate	Output size
Mapping layer	$1\times1\times109$	8	1	–	91×91
ConvBlock1-Conv	5×5	16	2	–	44×44
ConvBlock1-MaxP	2×2	–	2	–	22×22
ConvBlock2-Conv	3×3	32	1	–	20×20
ConvBlock2-MaxP	2×2	–	2	–	10×10
ConvBlock3-Conv	3×3	64	1	–	8×8
ConvBlock3-MaxP	2×2	–	2	–	4×4
ConvBlock4-Conv	3×3	128	1	–	2×2
ConvBlock4-MaxP	2×2	–	2	–	1×1
Flatten	–	–	–	–	128
FCBlock1	–	256	–	0.5	256
FCBlock2	–	64	–	0.5	64
FC3	–	2	–	–	2

into training, validation and testing sets, which account for 80%, 10% and 10% of the dataset for each task, respectively. The model is trained for 150 epochs and the best performing model with the lowest objective function value on the validation set is saved and its performance is evaluated on the testing set. This procedure is then repeated 10 times with different sampling seeds so as to have different samples in the train/validation/testing splits and minimize the effect of random variation.

6.3.2 Evaluation on single view CNN

The results obtained by single view CNN are presented in Table 6.3 and Table 6.4 for AD vs. NC and pMCI vs. sMCI, respectively. As can be seen, three views have nearly identical performance in AD diagnosis with slight differences. Coronal view outperforms the other two views in terms of SPE and AUC, which are 94.50% and 94.30% respectively, while sagittal view performs better than others concerning ACC (90.21%) and SEN (87.12%). For classifying pMCI from sMCI, basically coronal view can yield good performance compared to axial and sagittal

views since it achieves highest values concerning ACC, SPE and bACC, which are 74.91%, 82.07% and 72.12%, respectively. Due to the unbalanced dataset for pMCI vs. sMCI, the significance expressed by ACC has been weakened. Instead, AUC and bACC which derived from SEN and SPE can gain more weights in the evaluation. The metric AUC indicates axial view has an advantage, while bACC implies coronal view performs well. Considering differences between the coronal view and the axial view regarding AUC and bACC, coronal view can indeed yield good overall performance. Furthermore, one interesting thing is that despite of the highest AUC achieved by axial view, it also gives the lowest bACC, 69.47%. It is attributed to the character of AUC which is insensitive to unbalanced dataset. Generally, when computing accuracy, the obtained probability needs to be transformed to a label, for which a threshold should be set. Different thresholds will lead to different accuracies. Basically, the threshold is set to a default value, 0.5, which is suitable for balanced case, while for the unbalanced situation, 0.5 is obviously a biased value. Therefore, the balanced accuracy, 69.47%, achieved by axial view is a relatively biased result. From this point of view, AUC is the most convincing metric, but bACC has a good interpretability. In summary, there is no single view dominating the performance with a big advantage, thus three views should be considered jointly.

Table 6.3: Performance of single view for AD vs. NC (%)

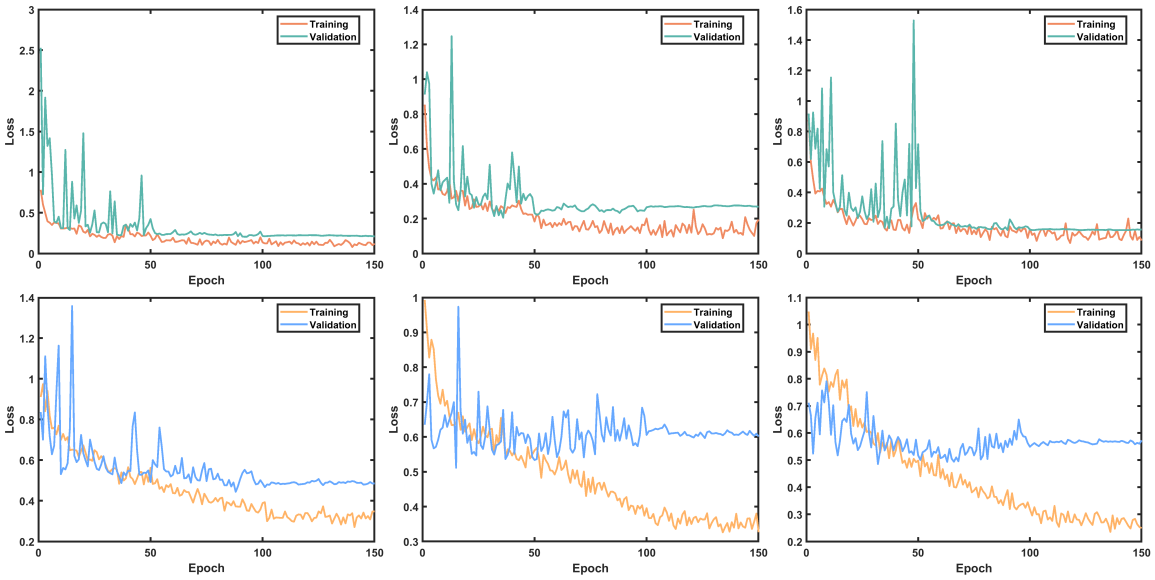
Single view	ACC	SEN	SPE	AUC
Axial view	88.33	83.71	92.64	94.02
Coronal view	89.38	84.48	**94.50**	**94.30**
Sagittal view	**90.21**	**87.12**	93.23	93.34

Table 6.4: Performance of single view for pMCI vs. sMCI (%)

Single view	ACC	SEN	SPE	AUC	bACC
Axial view	71.74	61.23	77.71	**77.23**	69.47
Coronal view	**74.91**	62.16	**82.07**	77.08	**72.12**
Sagittal view	74.03	**64.67**	79.16	76.66	71.92

The loss curves of training and validation sets are illustrated in Figure 6.6 in which the top row is for AD diagnosis and the bottom row is for pMCI vs. sMCI. It can be seen that loss curves tend to flat with the increase of epochs for all the three views, which implies that the model is converged for each view. The larger

differences between training and validation sets in coronal and sagittal views, especially obtained in the case of pMCI vs. sMCI, are caused by smaller probability differences between positive and negative labels.

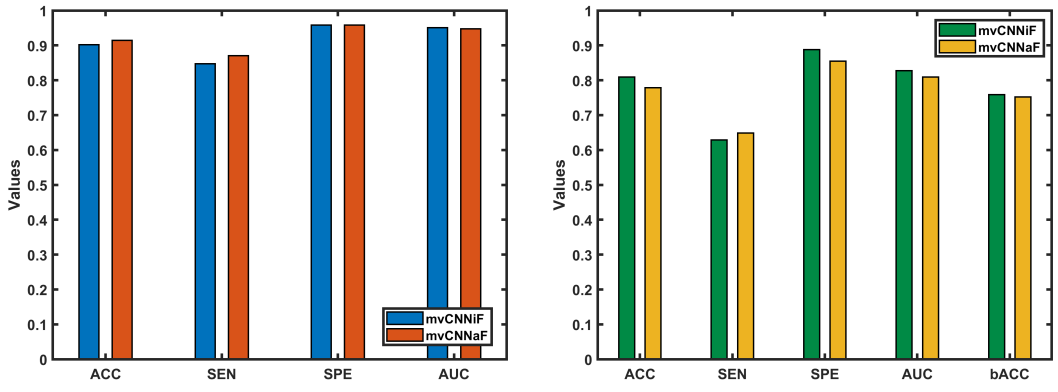


(a) (b) (c)

Figure 6.6: Training and validation losses of different views during a training procedure where the top row is for AD vs.NC and the bottom row is for pMCI vs. sMCI. (a) Axial view. (b) Coronal view. (c) Sagittal view.

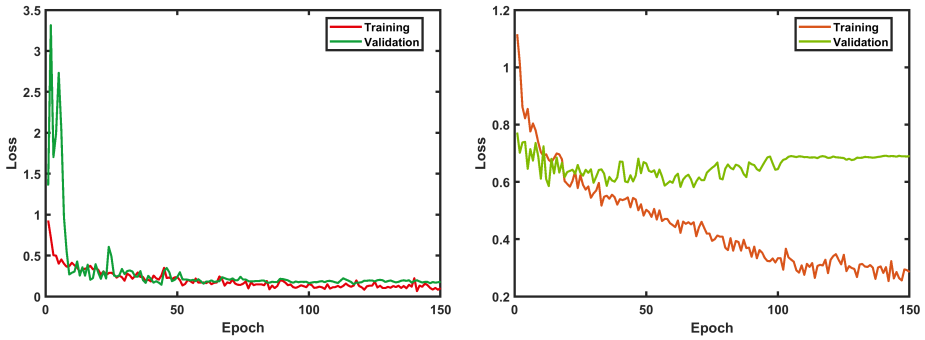
6.3.3 Evaluation on multiple view CNN

Figure 6.7 shows the performance of the two types of multiview CNN models, mvCNNiF and mvCNNaF, for AD vs. NC and pMCI vs. sMCI. As can be seen, the mvCNNiF model is slightly inferior to mvCNNaF in AD diagnosis with difference of 1.26%, 2.29%, −0.08% and −0.34% in terms of ACC, SEN, SPE and AUC, respectively, while mvCNNaF can achieve 91.46%, 87.04%, 95.78% and 94.77% regarding the four metrics. In contrast, mvCNNiF outperforms mvCNNaF in classifying pMCI from sMCI, which obtains 80.92%, 62.85%, 88.82%, 82.73% and 75.84% concerning ACC, SEN, SPE, AUC and bACC. Therefore, according to experiment results, mvCNNiF is suitable for pMCI vs. sMCI, while mvCNNaF works well for AD diagnosis. In addition, the loss curves of training and validation sets generated by mvCNNiF are shown in Figure 6.8 for AD vs. NC and pMCI vs. sMCI. It can be seen that mvCNNiF can get converged for two tasks.



(a) (b)

Figure 6.7: Performance of multiview CNN. (a) AD vs. NC. (b) pMCI vs. sMCI.



(a) (b)

Figure 6.8: Training and validation losses of mvCNNiF. (a) AD vs. NC (b) pMCI vs. sMCI.

6.3.4 Comparison with state-of-the-art methods

The three proposed methods, including multilevel feature representation, multiscale spatial gradient features and multiview CNN, are compared with other methods which also address the problem of AD/pMCI diagnosis under the modality of FDG-PET, including methods of Li *et al.* [47], Gray *et al.* [55], Padilla *et al.* [68], Hinrichs *et al.* [70], Zhu *et al.* [77], Lu *et al.* [114] and Liu *et al.* [116]. Among these comparison methods, Lu's and Liu's methods applied the neural network technique, and other approaches were developed under the traditional classification framework.

Table 6.5 and Table 6.6 present the comparison results with state-of-the-art methods for two binary tasks, AD vs. NC and pMCI vs. sMCI. It can be seen that the proposed method which utilizes multiscale spatial gradient features outperforms others in terms of ACC, SPE and AUC for AD diagnosis, but it is inferior to Lu's method in respect of SEN, with a slight difference of 0.09%. The other two pro-

posed methods, multilevel feature representation and multiview CNN (referred to mvCNNaF), also yield significant results which are higher than those obtained by most comparison methods. For the case of distinguishing between pMCI and sMCI, the multiview CNN (referred to mvCNNiF) outperforms the other two proposed methods. While the best method, mvCNNiF, is not as effective as Lu's method which utilized multiscale deep neural networks in terms of ACC, SEN and bACC with differences of 0.63%, 10.48% and 2.74%, respectively. It should be noted that the data used in Lu's method for the task of pMCI vs. sMCI is more unbalanced (pMCI: 112, sMCI: 409) than the dataset used in this thesis. Therefore, the metric AUC could give a more convincing evaluation, but it is not clarified in Lu's method.

Table 6.5: Performance comparison with state-of-the-art methods for AD vs. NC(%)

Method	Subjects	ACC	SEN	SPE	AUC
Li *et al.* [47]	25AD + 30NC	89.1	92	86	97
Gray *et al.* [55]	50AD + 54NC	88.4	83.2	93.6	−−
Padilla *et al.* [68]	53AD + 52NC	86.59	87.50	85.36	−−
Hinrichs *et al.* [70]	89AD + 94NC	84	84	82	87.16
Zhu *et al.* [77]	51AD + 52NC	93.3	−−	−−	−−
Lu *et al.* [114]	226AD + 304NC	93.58	**91.54**	95.06	−−
Liu *et al.* [116]	93AD + 100NC	91.2	91.4	91.0	95.3
Multilevel	237AD + 242NC	92.57	90.89	94.42	96.83
Multiscale	237AD + 242NC	**94.20**	91.45	**96.76**	**97.42**
mvCNNaF	237AD + 242NC	91.46	87.04	95.78	94.77

Table 6.6: Performance comparison with state-of-the-art methods for pMCI vs. sMCI(%)

Method	Subjects	ACC	SEN	SPE	AUC	bACC
Gray *et al.* [55]	53pMCI + 64sMCI	63.1	52.2	73.2	−−	62.7
Zhu *et al.* [77]	43pMCI + 56sMCI	69.9	−−	−−	−−	−−
Lu *et al.* [114]	112pMCI + 409sMCI	**81.55**	**73.33**	83.83	−−	**78.58**
Multilevel	209pMCI + 360sMCI	76.17	69.57	80.26	81.95	74.92
Multiscale	209pMCI + 360sMCI	76.85	67.94	82.43	82.10	75.18
mvCNNiF	209pMCI + 360sMCI	80.92	62.85	**88.82**	**82.73**	75.84

6.4 Conclusion

In this chapter, two types of CNN architectures are proposed to address the problems of AD diagnosis and MCI conversion prediction, mvCNNiF and mvCNNaF, which take axial, coronal and sagittal views into account jointly. Benefit from a proposed mapping layer which projects information along the third dimension onto a plane, both models can perform 2D convolution operations for each view and meanwhile consider the spatial information. Experimental results indicate that no single view has an obvious advantage in AD diagnosis or identifying pMCI from sMCI, and moreover, the proposed models can achieve significant performance, especially for mvCNNiF in MCI conversion prediction, which surpasses the multilevel feature representation and multiscale gradient methods.

Chapter 7

Conclusions and Perspectives

This thesis is devoted to addressing problems of AD diagnosis and MCI conversion prediction by using FDG-PET modality. Three approaches have been proposed.

In Chapter 4, a multilevel feature representation method is proposed which considers not only the anatomical regions' properties, but also connectivities among regions that are rarely taken into account by most methods. Then a similarity-driven ranking method is developed to reduce the feature dimension and increases the classifier's diversity. Last multiple levels of features are fed into multiple classifiers and selected classifiers are then integrated into an ensemble model to give a decision.

Chapter 5 attempts to further extend the feature pool of FDG-PET representation, in which FDG-PET images are characterized from the view of spatial gradients. Multiscale spatial gradient features, SSH and LSH, are included to an ensemble classifier to make a prediction jointly, which considers performance of individual and concatenated regions at the same time. Besides, attributed to a proposed region ranking method which involves multiple SSH features, the ensemble model can yield better performance by exploiting less regions.

In Chapter 6, two types of multiview CNN architectures, mvCNNiF and mvCNNaF, are developed so as to take axial, coronal and sagittal views into account simultaneously. A novel mapping layer is introduced to CNN models in order to project information along the third dimension onto a plane, which converts a 3D problem to a 2D problem, thereby not only reducing parameters involved in convolutional layers but also considering the spatial relations.

In order to evaluate the performance of proposed methods, experiments are conducted on a public dataset, ADNI dataset. In Chapter 3, the procedure of data acquisition and rules of data selection are clarified in order to provide guidance for

other methods. Accordingly, we have obtained totally 1048 FDG-PET images from the baseline scans of 1048 subjects, respectively. It is basically the most complete dataset, which can ensure the effectiveness of evaluations for proposed methods.

The CNN-based models are not as effective as the conventional machine learning methods in AD diagnosis, therefore hyperparameters involved in mvCNNiF or mvCNNaF should be further adjusted and tested so as to yield impressive results. Beyond methods proposed in this thesis, the future works mainly lie in three folds:

- Data

  Multiple modalities, especially for the combination of MRI and FDG-PET, can be utilized to tackle the problem of AD diagnosis. In addition, longitudinal data is also a significant topic, which can not only expand the dataset, but also provide an insight into the progression of AD. Therefore, longitudinal data should be investigated as well in the future.

- Methods

  We just explore the effectiveness of the first order deviations, spatial gradients, in Chapter 5, but the second order deviations are also worth investigating. Besides, other descriptors, such as LBP (Local Binary Pattern), SIFT (Scale Invariant Feature Transform), are required to test the corresponding performance so as to verify that those features derived from natural scene images are useful for neuroimaging data. As to deep learning models, Generative Adversarial Network (GAN) can be introduced to learn useful features automatically.

- Tasks

  AD diagnosis or MCI conversion prediction is the most commonly addressed topic. Other related tasks are also interesting, for instance, the prediction of clinical measures which could include mini–mental state examination (MMSE) and clinical dementia rating sum of boxes (CDR-SB), or localization of regions with low metabolism. Another significant issue could be prediction visualization. The current works mainly focus on the prediction results, diseased or not, converted or not. But if we could predict the progression and then generate its corresponding neuroimaging and visualize this procedure, it would be more intuitive to explain, probe or understand Alzheimer's Disease.

Bibliography

[1] N. C. Berchtold and C. W. Cotman. Evolution in the conceptualization of dementia and alzheimer's disease: Greco-roman period to the 1960s. *Neurobiology of Aging*, 19(3):173–189, 1998.

[2] C. Ballard, S. Gauthier, A. Corbett, C. Brayne, D. Aarsland, and E. Jones. Alzheimer's disease. *Lancet*, 377:1019–908.

[3] H. Förstl and A. Kurz. Clinical features of alzheimer's disease. *European archives of psychiatry and clinical neuroscience*, 249(6):288–290, 1999.

[4] V. Taler and N. A. Phillips. Language performance in alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of clinical and experimental neuropsychology*, 30(5):501–556, 2008.

[5] Association Alzheimer's. 2015 alzheimer's disease facts and figures. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 11(3):332, 2015.

[6] L. E. Hebert, J. Weuve, P. A. Scherr, and D. A. Evans. Alzheimer disease in the united states (2010–2050) estimated using the 2010 census. *Neurology*, 80(19):1778–1783, 2013.

[7] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou. World alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future. 2016.

[8] R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack Jr, J. Kaye, T. J. Montine, et al. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia*, 7(3):280–292, 2011.

[9] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, 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. *Alzheimer's & dementia*, 7(3):270–279, 2011.

[10] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack Jr, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, 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. *Alzheimer's & dementia*, 7(3):263–269, 2011.

[11] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan. Clinical diagnosis of alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, 34(7):939–939, 1984.

[12] D. C. Delis. *CVLT-C: California verbal learning test*. Psychological Corporation, Harcourt Brace Corporation, 1994.

[13] E. Kaplan, H. Googlass, and S. Weintrab. Boston naming test. Philadelphia: Lea and Febiger; 1983. *Dement Geriatr Cogn Disord*, 20:198–208, 2005.

[14] C. A. Lynch, C. Walsh, A. Blanco, M. Moran, R. F. Coen, J. B. Walsh, and B. A. Lawlor. The clinical dementia rating sum of box score in mild dementia. *Dementia and geriatric cognitive disorders*, 21(1):40–43, 2006.

[15] C. R Jack Jr, D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, and J. Q. Trojanowski. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

[16] R. Sperling. Functional MRI studies of associative encoding in normal aging, mild cognitive impairment, and alzheimer's disease. *Annals of the New York Academy of Sciences*, 1097(1):146–155, 2007.

[17] J. S. Damoiseaux. Resting-state fMRI as a biomarker for alzheimer's disease? *Alzheimer's research & therapy*, 4(2):8, 2012.

[18] S. Teipel, A. Drzezga, M. J. Grothe, H. Barthel, G. Chételat, N. Schuff, P. Skudlarski, E. Cavedo, G. B. Frisoni, W. Hoffmann, et al. Multimodal imaging in Alzheimer's disease: validity and usefulness for early detection. *The Lancet Neurology*, 14(10):1037–1053, 2015.

[19] P. Selnes, D. Aarsland, A. Bjørnerud, L. Gjerstad, A. Wallin, E. Hessen, I. Reinvang, R. Grambaite, V. K. Auning, E.and Kjærvik, et al. Diffusion tensor imaging surpasses cerebrospinal fluid as predictor of cognitive decline and medial temporal lobe atrophy in subjective cognitive impairment and mild cognitive impairment. *Journal of Alzheimer's disease*, 33(3):723–736, 2013.

[20] E. Scola, M. Bozzali, F. Agosta, G. Magnani, M. Franceschi, M. P. Sormani, M. Cercignani, E. Pagani, M. Falautano, M. Filippi, et al. A diffusion tensor MRI study of patients with MCI and AD with a 2-year clinical follow-up. *Journal of Neurology, Neurosurgery & Psychiatry*, 81(7):798–805, 2010.

[21] A. Fellgiebel, P. R. Dellani, D. Greverus, A. Scheurich, P. Stoeter, and M. J. Müller. Predicting conversion to dementia in mild cognitive impairment by volumetric and diffusivity measurements of the hippocampus. *Psychiatry Research: Neuroimaging*, 146(3):283–287, 2006.

[22] D. Le Bihan, J. F. Mangin, C. Poupon, C. A. Clark, S. Pappata, N. Molko, and H. Chabriat. Diffusion tensor imaging: concepts and applications. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 13(4):534–546, 2001.

[23] U. Saeed, W. Swardfager, S. E. Black, and M. Masellis. Biomarkers of Alzheimer's disease. *Mental Health and Illness of the Elderly*, pages 105–139, 2017.

[24] G. B. Saha, W. J. MacIntyre, and R. T. Go. Radiopharmaceuticals for brain imaging. *Seminars in nuclear medicine*, 24(4):324–349, 1994.

[25] R. Ceravolo, D. Volterrani, G. Gambaccini, C. Rossi, C. Logi, G. Manca, C. Berti, G.iuliano Mariani, L. Murri, and U. Bonuccelli. Dopaminergic degeneration and perfusional impairment in Lewy body dementia and Alzheimer's disease. *Neurological Sciences*, 24(3):162–163, 2003.

[26] S. J. Colloby, J. D. Fenwick, D. E. Williams, S. M. Paling, K. Lobotesis, C. Ballard, I. McKeith, and J. T. O'Brien. A comparison of 99m Tc-HMPAO

SPET changes in dementia with lewy bodies and Alzheimer's disease using statistical parametric mapping. *European journal of nuclear medicine and molecular imaging*, 29(5):615–622, 2002.

[27] W. Jagust, R. Thisted, M. D. Devous, R. Van Heertum, H. Mayberg, K. Jobst, A. D. Smith, and N. Borys. SPECT perfusion imaging in the diagnosis of Alzheimer's disease: a clinical-pathologic study. *Neurology*, 56(7):950–956, 2001.

[28] E. Guedj, E. J. Barbeau, M. Didic, O. Felician, C. De Laforte, M. Ceccaldi, O. Mundler, and M. Poncet. Identification of subgroups in amnestic mild cognitive impairment. *Neurology*, 67(2):356–358, 2006.

[29] D. L. Bailey, M. N. Maisey, D. W. Townsend, and P. E. Valk. *Positron emission tomography*. Springer, 2005.

[30] M. J. Pontecorvo and M. A. Mintun. PET amyloid imaging as a tool for early diagnosis and identifying patients at risk for progression to Alzheimer's disease. *Alzheimer's research & therapy*, 3(2):11, 2011.

[31] A. M. Catafau and S. Bullich. Amyloid PET imaging: applications beyond Alzheimer's disease. *Clinical and translational imaging*, 3(1):39–55, 2015.

[32] R. E. Coleman. Positron emission tomography diagnosis of Alzheimer's disease. *Neuroimaging Clinics*, 15(4):837–846, 2005.

[33] W. Jagust, B. Reed, D. Mungas, W. Ellis, and C. Decarli. What does fluorodeoxyglucose PET imaging add to a clinical diagnosis of dementia? *Neurology*, 69(9):871–877, 2007.

[34] A. Drzezga, T. Grimmer, M. Riemenschneider, N. Lautenschlager, H. Siebner, P. Alexopoulus, S. Minoshima, M. Schwaiger, and A. Kurz. Prediction of individual clinical outcome in MCI by means of genetic assessment and 18F-FDG PET. *Journal of Nuclear Medicine*, 46(10):1625–1632, 2005.

[35] R. KJ. Brown, N. I. Bohnen, K. K. Wong, S. Minoshima, and K. A. Frey. Brain PET in suspected dementia: patterns of altered FDG metabolism. *Radiographics*, 34(3):684–701, 2014.

[36] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society, 2005.

[37] Rik Vandenberghe, Natalie Nelissen, Eric Salmon, Adrian Ivanoiu, Steen Hasselbalch, Allan Andersen, Alex Korner, Lennart Minthon, David J Brooks, Koen Van Laere, et al. Binary classification of 18f-flutemetamol pet using machine learning: comparison with visual reads and structural mri. *NeuroImage*, 64:517–525, 2013.

[38] Katherine R Gray, Paul Aljabar, Rolf A Heckemann, Alexander Hammers, Daniel Rueckert, Alzheimer's Disease Neuroimaging Initiative, et al. Random forest-based similarity measures for multi-modal classification of alzheimer's disease. *NeuroImage*, 65:167–175, 2013.

[39] Tong Tong, Katherine Gray, Qinquan Gao, Liang Chen, Daniel Rueckert, Alzheimer's Disease Neuroimaging Initiative, et al. Multi-modal classification of alzheimer's disease using nonlinear graph fusion. *Pattern recognition*, 63:171–181, 2017.

[40] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

[41] Edmund T Rolls, Marc Joliot, and Nathalie Tzourio-Mazoyer. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *Neuroimage*, 122:1–5, 2015.

[42] David W Shattuck, Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L Narr, Russell A Poldrack, Robert M Bilder, and Arthur W Toga. Construction of a 3d probabilistic atlas of human cortical structures. *Neuroimage*, 39(3):1064–1080, 2008.

[43] Alexander Hammers, Richard Allom, Matthias J Koepp, Samantha L Free, Ralph Myers, Louis Lemieux, Tejal N Mitchell, David J Brooks, and John S Duncan. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Human brain mapping*, 19(4):224–247, 2003.

[44] Ioannis S Gousias, Daniel Rueckert, Rolf A Heckemann, Leigh E Dyet, James P Boardman, A David Edwards, and Alexander Hammers. Automatic segmentation of brain mris of 2-year-olds into 83 regions of interest. *Neuroimage*, 40(2):672–684, 2008.

[45] Rolf A Heckemann, Joseph V Hajnal, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. Automatic anatomical brain mri segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126, 2006.

[46] Isabelle Faillenot, Rolf A Heckemann, Maud Frot, and Alexander Hammers. Macroanatomy and 3d probabilistic atlas of the human insula. *NeuroImage*, 150:88–98, 2017.

[47] Rui Li, Robert Perneczky, Igor Yakushev, Stefan Foerster, Alexander Kurz, Alexander Drzezga, Stefan Kramer, Alzheimer's Disease Neuroimaging Initiative, et al. Gaussian mixture models and model selection for [18f] fluorodeoxyglucose positron emission tomography classification in alzheimer's disease. *PloS one*, 10(4):e0122731, 2015.

[48] Jorge Samper-Gonzalez, Ninon Burgos, Simona Bottani, Sabrina Fontanella, Pascal Lu, Arnaud Marcoux, Alexandre Routier, Jérémy Guillon, Michael Bacci, Junhao Wen, et al. Reproducible evaluation of classification methods in alzheimer's disease: Framework and application to mri and pet data. *NeuroImage*, 183:504–521, 2018.

[49] M Pagani, F De Carli, S Morbelli, J Öberg, A Chincarini, GB Frisoni, S Galluzzi, R Perneczky, A Drzezga, BNM Van Berckel, et al. Volume of interest-based [18f] fluorodeoxyglucose pet discriminates mci converting to alzheimer's disease from healthy controls. a european alzheimer's disease consortium (eadc) study. *NeuroImage: Clinical*, 7:34–42, 2015.

[50] Marco Pagani, Flavio Nobili, Silvia Morbelli, Dario Arnaldi, Alessandro Giuliani, Johanna Öberg, Nicola Girtler, Andrea Brugnolo, Agnese Picco, Matteo Bauckneht, et al. Early identification of mci converting to ad: a fdg pet study. *European journal of nuclear medicine and molecular imaging*, 44(12):2042–2052, 2017.

[51] Marco Pagani, Alessandro Giuliani, Johanna Öberg, Andrea Chincarini, Silvia Morbelli, Andrea Brugnolo, Dario Arnaldi, Agnese Picco, Matteo Bauckneht,

Ambra Buschiazzo, et al. Predicting the transition from normal aging to alzheimer's disease: a statistical mechanistic evaluation of FDG-PET data. *Neuroimage*, 141:282–290, 2016.

[52] Imene Garali, Mouloud Adel, Salah Bourennane, and Eric Guedj. Brain region ranking for 18fdg-pet computer-aided diagnosis of alzheimer's disease. *Biomedical Signal Processing and Control*, 27:15–23, 2016.

[53] Imène Garali, Mouloud Adel, Salah Bourennane, and Eric Guedj. Histogram-based features selection and volume of interest ranking for brain pet image classification. *IEEE journal of translational engineering in health and medicine*, 6:1–12, 2018.

[54] Yupeng Li, Jiehui Jiang, Jiaying Lu, Juanjuan Jiang, Huiwei Zhang, and Chuantao Zuo. Radiomics: a novel feature extraction method for brain neuron degeneration disease using 18f-fdg pet imaging and its implementation for alzheimer's disease and mild cognitive impairment. *Therapeutic Advances in Neurological Disorders*, 12:1756286419838682, 2019.

[55] K. R. Gray, R. Wolz, R. A. Heckemann, P. Aljabar, A. Hammers, D. Rueckert, Alzheimer's Disease Neuroimaging Initiative, et al. Multi-region analysis of longitudinal fdg-pet for the classification of alzheimer's disease. *NeuroImage*, 60(1):221–229, 2012.

[56] Kenichi Ota, Naoya Oishi, Kengo Ito, Hidenao Fukuyama, Sead-J Study Group, Alzheimer's Disease Neuroimaging Initiative, et al. Effects of imaging modalities, brain atlases and feature selection on prediction of alzheimer's disease. *Journal of neuroscience methods*, 256:168–183, 2015.

[57] Yousra Asim, Basit Raza, Ahmad Kamran Malik, Saima Rathore, Lal Hussain, and Mohammad Aksam Iftikhar. A multi-modal, multi-atlas-based approach for alzheimer detection via machine learning. *International Journal of Imaging Systems and Technology*, 28(2):113–123, 2018.

[58] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[59] Joseph Lee Rodgers and W Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[60] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (8):1226–1238, 2005.

[61] Stefan J Teipel, Jens Kurth, Bernd Krause, Michel J Grothe, Alzheimer's Disease Neuroimaging Initiative, et al. The relative importance of imaging markers for the prediction of alzheimer's disease dementia in mild cognitive impairment—beyond classical regression. *NeuroImage: Clinical*, 8:583–593, 2015.

[62] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[63] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[64] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[65] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[66] Peter A Lachenbruch and M Goldstein. Discriminant analysis. *Biometrics*, pages 69–85, 1979.

[67] Aleix M Martínez and Avinash C Kak. Pca versus lda. *IEEE transactions on pattern analysis and machine intelligence*, 23(2):228–233, 2001.

[68] Pablo Padilla, Miriam López, Juan Manuel Górriz, Javier Ramirez, Diego Salas-Gonzalez, and I Alvarez. Nmf-svm based cad tool applied to functional brain images for the diagnosis of alzheimer's disease. *IEEE Transactions on medical imaging*, 31(2):207–216, 2012.

[69] D Salas-Gonzalez, JM Górriz, J Ramírez, IA Illán, M López, F Segovia, R Chaves, P Padilla, CG Puntonet, and Alzheimer's Disease Neuroimage Initiative. Feature selection using factor analysis for alzheimer's diagnosis using pet images. *Medical physics*, 37(11):6084–6095, 2010.

[70] Chris Hinrichs, Vikas Singh, Lopamudra Mukherjee, Guofan Xu, Moo K Chung, Sterling C Johnson, Alzheimer's Disease Neuroimaging Initiative, et al. Spatially augmented lpboosting for ad classification with evaluations on the adni dataset. *Neuroimage*, 48(1):138–149, 2009.

[71] Carlos Cabral, Pedro M Morgado, Durval Campos Costa, Margarida Silveira, Alzheimer's Disease Neuroimaging Initiative, et al. Predicting conversion from mci to ad with FDG-PET brain images at different prodromal stages. *Computers in biology and medicine*, 58:101–109, 2015.

[72] IA Illán, J Manuel Górriz, Javier Ramírez, Diego Salas-Gonzalez, MM López, Fermín Segovia, Rosa Chaves, Manuel Gómez-Rio, Carlos García Puntonet, Alzheimer's Disease Neuroimaging Initiative, et al. 18f-fdg pet imaging analysis for computer aided alzheimer's diagnosis. *Information Sciences*, 181(4):903–916, 2011.

[73] Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Multimodal classification of alzheimer's disease and mild cognitive impairment. *Neuroimage*, 55(3):856–867, 2011.

[74] Noor Jehan Kabani, David J MacDonald, Colin J Holmes, and Alan C Evans. 3d anatomical atlas of the human brain. *NeuroImage*, 7(4):S717, 1998.

[75] Yinghuan Shi, Heung-Il Suk, Yang Gao, and Dinggang Shen. Joint coupled-feature representation and coupled boosting for ad diagnosis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2721–2728, 2014.

[76] Chen Zu, Biao Jie, Mingxia Liu, Songcan Chen, Dinggang Shen, Daoqiang Zhang, Alzheimer's Disease Neuroimaging Initiative, et al. Label-aligned multi-task feature learning for multimodal classification of alzheimer's disease and mild cognitive impairment. *Brain imaging and behavior*, 10(4):1148–1159, 2016.

[77] Xiaofeng Zhu, Heung-Il Suk, Seong-Whan Lee, and Dinggang Shen. Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification. *IEEE Transactions on Biomedical Engineering*, 63(3):607–618, 2015.

[78] Andrew Y Ng and Michael I Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848, 2002.

[79] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[80] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[81] J Ross Quinlan. *C4.5: programs for machine learning*. Elsevier, 2014.

[82] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[83] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[84] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[85] Paul Geladi and Bruce R Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17, 1986.

[86] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25, 1982.

[87] John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.

[88] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[89] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC, 2012.

[90] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[91] Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

[92] Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[93] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[94] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[95] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[96] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[97] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124. IEEE, 2010.

[98] Kent A Spackman. Signal detection theory: Valuable tools for evaluating inductive learning. In *Proceedings of the sixth international workshop on Machine learning*, pages 160–163. Elsevier, 1989.

[99] Geisser Seymour. Predictive inference. *0412034719Chapman and Hall, New York*, 1993.

[100] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[101] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[102] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[103] Rui Min, Guorong Wu, Jian Cheng, Qian Wang, Dinggang Shen, and Alzheimer's Disease Neuroimaging Initiative. Multi-atlas based representations for alzheimer's disease diagnosis. *Human brain mapping*, 35(10):5052–5070, 2014.

[104] Christiane Möller, Yolande AL Pijnenburg, Wiesje M van der Flier, Adriaan Versteeg, Betty Tijms, Jan C de Munck, Anne Hafkemeijer, Serge ARB Rombouts, Jeroen van der Grond, John van Swieten, et al. Alzheimer disease and behavioral variant frontotemporal dementia: automatic classification based on cortical atrophy for single-subject diagnosis. *Radiology*, 279(3):838–848, 2015.

[105] Mingxia Liu, Daoqiang Zhang, and Dinggang Shen. Relationship induced multi-template learning for diagnosis of alzheimer's disease and mild cognitive impairment. *IEEE transactions on medical imaging*, 35(6):1463–1474, 2016.

[106] Biao Jie, Daoqiang Zhang, Wei Gao, Qian Wang, Chong-Yaw Wee, and Dinggang Shen. Integration of network topological and connectivity properties for neuroimaging classification. *IEEE transactions on biomedical engineering*, 61(2):576–589, 2014.

[107] Ali Khazaee, Ata Ebrahimzadeh, and Abbas Babajani-Feremi. Identifying patients with alzheimer's disease using resting-state fmri and graph theory. *Clinical Neurophysiology*, 126(11):2132–2141, 2015.

[108] Talia M Nir, Julio E Villalon-Reina, Gautam Prasad, Neda Jahanshad, Shantanu H Joshi, Arthur W Toga, Matt A Bernstein, Clifford R Jack Jr, Michael W Weiner, Paul M Thompson, et al. Diffusion weighted imaging-based maximum density path analysis and classification of alzheimer's disease. *Neurobiology of aging*, 36:S132–S140, 2015.

[109] Gautam Prasad, Shantanu H Joshi, Talia M Nir, Arthur W Toga, Paul M Thompson, Alzheimer's Disease Neuroimaging Initiative (ADNI, et al. Brain connectivity and novel network measures for alzheimer's disease classification. *Neurobiology of aging*, 36:S121–S131, 2015.

[110] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[111] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[112] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

[113] Tao Zhou, Kim-Han Thung, Xiaofeng Zhu, and Dinggang Shen. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. *Human brain mapping*, 40(3):1001–1016, 2019.

[114] Donghuan Lu, Karteek Popuri, Gavin Weiguang Ding, Rakesh Balachandar, Mirza Faisal Beg, Alzheimer's Disease Neuroimaging Initiative, et al. Multiscale deep neural network based analysis of fdg-pet images for the early diagnosis of alzheimer's disease. *Medical image analysis*, 46:26–34, 2018.

[115] Siqi Liu, Sidong Liu, Weidong Cai, Hangyu Che, Sonia Pujol, Ron Kikinis, Dagan Feng, Michael J Fulham, et al. Multimodal neuroimaging feature learning for multiclass diagnosis of alzheimer's disease. *IEEE Transactions on Biomedical Engineering*, 62(4):1132–1140, 2014.

[116] Manhua Liu, Danni Cheng, and Weiwu Yan. Classification of alzheimer's disease by combination of convolutional and recurrent neural networks using fdg-pet images. *Frontiers in neuroinformatics*, 12:35, 2018.

[117] Ashish Gupta, Murat Ayhan, and Anthony Maida. Natural image bases to represent neuroimaging data. In *International conference on machine learning*, pages 987–994, 2013.

[118] Yiming Ding, Jae Ho Sohn, Michael G Kawczynski, Hari Trivedi, Roy Harnish, Nathaniel W Jenkins, Dmytro Lituiev, Timothy P Copeland, Mariam S Aboian, Carina Mari Aparici, et al. A deep learning model to predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain. *Radiology*, 290(2):456–464, 2018.

[119] Yechong Huang, Jiahang Xu, Yuncheng Zhou, Tong Tong, Xiahai Zhuang, et al. Diagnosis of alzheimer's disease via multi-modality 3d convolutional neural network. *arXiv preprint arXiv:1902.09904*, 2019.

[120] Simeon Spasov, Luca Passamonti, Andrea Duggento, Pietro Lio, Nicola Toschi, Alzheimer's Disease Neuroimaging Initiative, et al. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease. *Neuroimage*, 189:276–287, 2019.

[121] Ehsan Hosseini-Asl, Robert Keynton, and Ayman El-Baz. Alzheimer's disease diagnostics by adaptation of 3d convolutional network. In *2016 IEEE*

International Conference on Image Processing (ICIP), pages 126–130. IEEE, 2016.

[122] 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. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[123] Fan Li, Manhua Liu, Alzheimer's Disease Neuroimaging Initiative, et al. Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks. *Computerized Medical Imaging and Graphics*, 70:101–110, 2018.

[124] W. D Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[125] Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: a network visualization tool for human brain connectomics. *PloS one*, 8(7):e68910, 2013.

[126] Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

[127] Katelyn L Arnemann, Franziska Stöber, Sharada Narayan, Gil D Rabinovici, and William J Jagust. Metabolic brain networks in aging and preclinical alzheimer's disease. *NeuroImage: Clinical*, 17:987–999, 2018.

[128] Jinyong Chung, Kwangsun Yoo, Eunjoo Kim, Duk L Na, and Yong Jeong. Glucose metabolic brain networks in early-onset vs. late-onset alzheimer's disease. *Frontiers in aging neuroscience*, 8:159, 2016.

[129] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[130] Robert Gilmore Pontius Jr and Marco Millones. Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32(15):4407–4429, 2011.

[131] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[132] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[133] François Chollet et al. Keras, 2015.

[134] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[135] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[136] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.