BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

DEEP LEARNING FOR ALZHEIMER'S DISEASE: TOWARDS THE DEVELOPMENT OF AN ASSISTIVE DIAGNOSTIC TOOL

by

SHANGRAN QIU

B.A., Xi'an Jiaotong University, 2016

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
2022

© 2022 by
SHANGRAN QIU
All rights reserved

Approved by

First Reader
Vijaya B. Kolachalama, Ph.D.
Associate Professor of Medicine & Computer Science

Second Reader
Andrei E. Ruckenstein, Ph.D.
Professor of Physics

Third Reader
Kirill S. Korolev, Ph.D.
Associate Professor of Physics

Dedication

I would like to dedicate this dissertation to all frontline workers who fight against Alzheimer's disease and other dementias.

Acknowledgements

First and foremost, I am deeply indebted to my PhD advisor, Prof. Vijaya Kolachalama, for offering me this research opportunity and leading me to become a better scientist in the field of artificial intelligence in healthcare. I still remember the first conversation I had with Vijaya, when I was a second-year PhD student from the Department of Physics, and it was also not long after he joined Boston University School of Medicine as a new faculty member. Though we had very different academic backgrounds, he still put his trust in me and gave me enough time to learn how to do research in medicine. Without his inexhaustible support and continuous guidance during the past five years, it would not have been possible for me to complete my PhD study. One thing I deeply admire about Vijaya is his eagerness to project our work to real-world applications and his vision on impacting the future of medicine with cutting-edge research. I learnt from him the importance of appreciating the translational impact of these research projects. His high academic standards also pushed me away from my comfort zone and motivated me to accomplish several important milestones.

I would like to express my deepest appreciation to the members of my dissertation committee, Prof. Andrei Ruckenstein, Prof. Claudio Chamon, Prof. Rhoda Au and Prof. Kirill Korolev. They have been extremely supportive of my research, and I would not have reached this stage without their constructive insights and invaluable guidance.

It has been a privilege to work with my lab colleagues, as they always inspire me with unique insights and allow me to think from different perspectives. I have benefited exceedingly from collaborating with my lab mates. I learnt so much domain knowledge in

medicine from Matthew Miller, Prajakta Joshi, Joyce Lee, and Michael Romano who have expertise in neurology, radiology, and pathology. They also taught me how to translate machine learning results to clinically applicable findings. I also acquired a lot of help from Chonghua Xue, Xiao Zhou, Xiao Wang, Yi Zheng, and Gary Chang whenever I needed to tackle technical challenges. I also want to give my thanks to other lab members: Meagan Lauber, Diana Zhang, Diala Lteif, Mona Jalal, Caitlin Newman, Lindsey Claus, and Sandeep Sreerama for many beneficial discussions. It has been a great pleasure for me to work in such a collaborative team environment.

Finally, I sincerely want to express my appreciation to everyone who supported me over the years. I am truly lucky to have many wonderful friends: Ang Liu, Jingsong Shang, Zhuofa Chen, Jun Cao, Weijun Luo, Jilin Wang, Shan Huang, and Teng Gao, without whom I would not have had such an enjoyable life in Boston. Thanks also goes to my girlfriend Baiyu Zhang for her company over these years. Lastly and most importantly, I want to express special gratitude to my parents for their love and support. Without them, this day would not have been possible.

DEEP LEARNING FOR ALZHEIMER’S DISEASE: TOWARDS THE DEVELOPMENT OF AN ASSISTIVE DIAGNOSTIC TOOL

SHANGRAN QIU

Graduate School of Arts and Sciences, 2022

Major Professor: Vijaya B. Kolachalama, Associate Professor of Medicine & Computer Science

Abstract

The past decade has witnessed rapid advances at the intersection of machine learning and medicine. Owing to the tremendous amount of digitized hospital data, machine learning is poised to bring innovation to the traditional healthcare workflow. Though machine learning models have strong predictive power, it is challenging to translate a research project into a clinical tool partly due to the lack of a rigorous validation framework. In this dissertation, I presented a range of machine learning models that were trained to classify Alzheimer’s disease - a condition with an insidious onset - using routinely collected clinical data. In addition to reporting the model performance, I discussed several considerations, including feature selection, data harmonization, effect of confounding variables, diagnostic scope, model interpretability and validation, which are critical to the design, development, and validation of machine learning models. From the methodological standpoint, I presented a multidisciplinary collaboration in which medical domain knowledge which was obtained from experts and tissue examinations was tightly integrated with the interpretable outcomes derived from our machine learning frameworks. I demonstrated that the model, which generalized well on multiple independent cohorts, achieved diagnostic performance on par with a group of medical professionals. The interpretable analysis of our model showed that its underlying decision logic corresponds with expert ratings and neuropathological findings. Taken together, this work presented a machine learning system for classification of Alzheimer’s disease, marking an important milestone towards a translatable clinical application in the future.

Table of Content

Dedication iv

Acknowledgements v

Abstract vii

Table of Content ix

List of Tables xiii

List of Figures xvi

List of Abbreviations xxiii

Chapter 1 1

Introduction 1

- 1.1 Alzheimer's Disease and Dementia 1
- 1.2 Machine Learning and Deep Learning 4
- 1.3 Interpretable Machine Learning 6

Chapter 2 9

Machine Learning and Deep Learning 9

- 2.1 Traditional Machine Learning Models 9
- 2.2 Artificial Neural Network 9
- 2.3 Convolutional Neural Network 11
- 2.4 Optimization and Gradient Descent 13
- 2.5 Model Training and Evaluation 15
  - 2.5.1 Splitting the Data 15
  - 2.5.2 Overfitting and Regularization 16

2.5.3 Evaluation 17

Chapter 3 19

Interpretable Machine Learning 19

3.1 Introduction 19

3.2 Intrinsic and Post hoc Interpretability 21

3.2.1 Intrinsically Interpretable Models 21

3.2.2 Post hoc Interpretable Methods 23

3.3 Global Feature Importance 26

3.4 Saliency Approach 26

Chapter 4 31

Interpretable Deep Learning for Alzheimer’s Disease 31

4.1 Introduction 31

4.2 Dataset 33

4.2.1 Dataset Collection 33

4.2.2 Data Harmonization 35

4.3 Model Development 36

4.3.1 Saliency Method 36

4.3.2 Disease Probability Map 39

4.3.3 Multimodal Data Integration 39

4.4 Validation 41

4.4.1 Local Accuracy 41

4.4.2 Neuropathological Validation 43

4.4.3 Global Accuracy Against Neurologists 45
4.5 Conclusion 49
Chapter 5 51
Expert-level Deep Learning for Dementia Assessment 51
5.1 Introduction 51
5.2 Dataset 54
5.2.1 Dataset Collection and Feature Selection 54
5.2.2 Harmonization of MRI Scans 58
5.2.3 Harmonization of Non-imaging Features 61
5.2.4 Effect of Confounding Factors 63
5.3 Model development 64
5.3.1 Data Split and Model Architecture 64
5.3.2 Model Performance 69
5.4 SHAP-based Interpretability Analysis 72
5.4.1 Voxel-level Saliency Map 72
5.4.2 Feature Importance Ranking 77
5.5 Validation 80
5.5.1 Expert-level Validation 80
5.5.2 Neuropathological Validation 82
5.6 Conclusion 84
Chapter 6 90
Summary and Future Directions 90

Appendix 96

Appendix A. Diagnostic Criteria by Cohort 96

Appendix B. Feature Missing Rate 103

Appendix C. Classification Performance between AD and stratified nADD 105

Appendix D. MRI Model's ROC & PR Curves on 7 Cohorts 107

Appendix E. Network Node Definition 108

Appendix F. Model Performance with Different Feature Combinations 111

Appendix G. Model vs Experts Performance 113

Appendix H. Classification Performance from Simple Thresholding 116

Bibliography 118

Curriculum vitae 128

List of Tables

Table 4.1. Study population and characteristics. Four independent datasets were used for this study including: the ADNI dataset, the AIBL, the FHS, and the NACC. The ADNI dataset was randomly split in the ratio of 3:1:1, where 60% of it was used for model training, 20% of the data was used for internal validation and the rest was used for internal testing. The best performing model on the validation dataset was selected for making predictions on the ADNI test data as well as on the AIBL, FHS and NACC datasets, which served as external test datasets for model validation. All the MRI scans considered for this study were performed on individuals within ±6 months from the date of clinical diagnosis. AD = Alzheimer's disease; NA = not available; NC = normal cognition. 34

Table 5.1. Study population and characteristics. Eight independent datasets were used for this study, including NACC, ADNI, AIBL, FHS, LBDSU, NIFD, OASIS, and PPMI. The NACC dataset was used to develop three separate types of models: i) an MRI-only CNN model that exclusively utilized imaging data, ii) non-imaging models in the form of traditional machine learning classifiers, which did not use any MRI data, and iii) a fusion model that combined imaging and non-imaging data within a hybrid architecture joining a CNN to a CatBoost model. The MRI-only model was validated across all eight cohorts, whereas external validation of non-imaging and fusion models was performed only on the OASIS cohort. All the MRI scans considered for this study were performed on individuals within ±6 months from the date of clinical diagnosis. The p-value for each dataset indicates the statistical significance of inter-group differences per column. We used two-tailed ANOVA and $\chi2$ tests for continuous and categorical variables, respectively. NC = normal cognition, MCI = mild cognitive impairment, AD = Alzheimer's disease dementia, nADD = non-Alzheimer's disease dementia; NA = not available. *FHS education code: 0=high school did not graduate, 1=high school graduate, 2=some college graduate, 3=college graduate. The symbol ^ indicates that data was not available for some subjects. 56

Table E.1. Definition of network nodes. We presented the network visualization of the inter-region correlation structure of the brain from the axial and sagittal views. The node was defined to represent a particular region from the parcellated brain and the edges between nodes demonstrated the sign and degree of correlation between a pair of nodes. We parcellated the brain MRI into 95 regions using the segmentation mask provided from the Hammersmith Adult Brain Atlas. To project the 3D structure into an axial and sagittal plane, we redefined the node for best visualization purpose. The "index" and "Adult brain atlas" columns show the completed 95 structures and their corresponding indexes from the Hammersmith Adult Brain Atlas. The "sagittal view" and "axial view" columns demonstrate how we merged and re-indexed regions and the node index number is what we labeled each node. In the sagittal view, we focused on visualizing the correlation between the temporal lobe, frontal lobe, parietal lobe, occipital lobe, cerebellum, and brainstem. Specifically, we merged the same structures from the left and right hemisphere as a single node in the sagittal projection, thus ending up with a total of 33 final nodes as defined in this table. In the axial view, we excluded some of the structures that have been already shown in the sagittal view, for example, insula, the third ventricle, etc. The focus of the axial view is to reveal the correlation between cerebrum structures from the left and right hemispheres. Our selection of the axial nodes yielded 57 regions.......... 110

Table G.1. Comparison of model performance with the neurologists. We randomly sampled 100 subjects from the NACC dataset. For each of the selected subject, we provided MRI scan along with a set of non-imaging features as specified in the supplementary material to 17 neurologists for them to review and make a prediction on one of the 4 possible categories, i.e., normal cognition (NC), mild cognitive impairment (MCI), Alzheimer's disease (AD), non-AD dementia (nADD). To make a head-to-head comparison, we also tested our MRI model, non-imaging model and fusion model on the same 100 selected subjects. We reported performance metrics, including accuracy, F-1, sensitivity, specificity, and Matthew correlation coefficient (MCC) for various tasks as indicated by column names based on the predictions from (a) 17 neurologists, (b) the MRI model, (c) the non-imaging model and (d) the fusion model. The mean and standard deviation (std) from table (a) was calculated over all 17 neurologists, and the mean and std from the other tables was derived from 5-fold validation experiments. More specifically, the COG represents the full classification of NC, MCI, and DE cases). In addition, we reported the performance of binary classification of NC vs. non-NC ("COG_{NC}" column), MCI vs. non-MCI ("COG_{MCI}" column) and DE vs. non-DE ("COG_{DE}" column). We also reported the model's performance in detecting AD from the dementied subjects within the "ADD columns. Lastly, we reported the 4-way classification of NC, MCI, AD, nADD ("4-way" column)... 114

Table G.2. Comparison of model performance with the neuroradiologists. We randomly sampled 50 subjects from the NACC dataset. For each of the selected subject, we provided MRI scan along with a set of non-imaging features as specified in the supplementary material to 7 neuroradiologists for them to independently review and make a prediction on one of the 2 possible categories, i.e., Alzheimer's disease (AD) and non-AD dementia (nADD). We reported performance metrics, including accuracy, F-1, sensitivity, specificity, and Matthew correlation coefficient (MCC) for this binary classification task by considering AD as positive samples............ 115

Table H.8. Classification performance of each standalone neuropsychological test. To compare our machine learning models to sample thresholding of common neuropsychiatric tests, we measured the area under the curve (AUC) of the ROC curves, and averaged precision (AP) of the precision-recall curve for the COG_{NC}, the COG_{DE} and the ADD tasks, respectively on the NACC cohort. The AUC and AP were derived by simply thresholding on each of the raw neuropsychiatric test score. 117

List of Figures

Figure 2.1. Structure of a perceptron model. 10

Figure 2.2. Architecture of a convolutional neural network. 12

Figure 2.3. Splitting data for cross-validation. 18

Figure 3.1. A decision tree. 22

Figure 4.1. Schematic of the deep learning framework. The FCN model was developed using a patch-based strategy in which randomly selected samples of T1-weighted full MRI volumes were passed to the model for training (Step 1). The corresponding Alzheimer's disease status of the individual served as the output for the classification model. Given that the operation of FCNs is independent of input data size, the model led to the generation of participant-specific disease probability maps of the brain (Step 2). Selected voxels of high-risk from the disease probability maps were then passed to the MLP for binary classification of disease status (Model A in Step 3; MRI model). As a further control, we used only the non-imaging features including age, gender and MMSE and developed an MLP model (Model B in Step 3; non-imaging model). We also developed another model that integrated multimodal input data including the selected voxels of high-risk disease probability maps alongside age, gender and MMSE score to perform binary classification of Alzheimer's disease status (Model C in Step 3; Fusion model). 38

Figure 4.2. Subject-specific disease probability maps. (A) Disease probability maps generated by the FCN model highlight high-risk brain regions that are associated with AD pathology. Individual cases are shown where the blue color indicates low-risk and red indicates high-risk of AD. The first two individuals had a clinical diagnosis of AD whereas the other two were clinically confirmed to have normal cognition. (B–D) Axial, coronal, and sagittal stacks of disease probability maps from a single subject with clinically confirmed Alzheimer's disease are shown. All imaging planes were used to construct 3D disease probability maps. Red color indicates locally inferred probability of Alzheimer's disease > 0.5, whereas blue indicates < 0.5. 40

Figure 4.3. Local accuracy of the FCN model performance. (A) Voxel-wise maps of MCC were computed independently across all the datasets to demonstrate predictive performance derived from all regions within the brain. (B–D) Axial, coronal, and sagittal stacks of the MCC maps at each cross-section from a single subject, are shown. These maps were generated by averaging the MCC values on the ADNI test data. 42

Figure 4.4. Correlation of model findings with neuropathology. (A) Overlap of model predicted regions of high Alzheimer's disease risk with post-mortem findings of Alzheimer's disease pathology in a single subject. This subject had clinically confirmed Alzheimer's disease with affected regions including the bilateral asymmetrical temporal lobes and the right-side hippocampus, the cingulate cortex, the corpus callosum, part of the parietal lobe and the frontal lobe. The first column (i) shows MRI slices in three different planes followed by a column (ii), which shows corresponding model predicted disease probability maps. A cut-off value of 0.7 was chosen to delineate the regions of high Alzheimer's disease risk and overlapped with the MRI scan in the next column (iii). The next column (iv) depicts a segmented mask of cortical and subcortical structures of the brain obtained from FreeSurfer. A sequential color-coding scheme denotes different levels of pathology ranging from green (0, low) to pale red (4, high). The final column (v) shows the overlay of the magnetic resonance scan, disease probability maps of high Alzheimer's disease risk and the color-coded regions based on pathology grade. (B) We then qualitatively assessed trends of neuropathological findings from the FHS dataset (n = 11). The same color-coding scheme as described above was used to represent the pathology grade (0–4) in the heat maps. The boxes colored in 'white' in the heat maps indicate missing data. Using the Spearman's Rank correlation coefficient test, an increasing Alzheimer's disease probability risk was associated with a higher grade of amyloid-β and tau accumulation, in the hippocampal formation, the middle frontal region, the amygdala and the temporal region, respectively. 44

Figure 4.5. Performance of the MLP model for Alzheimer's disease classification and model comparison with neurologists. (A) Sensitivity-specificity and precision-recall curves showing the sensitivity, the true positive rate, versus specificity, the true negative rate, calculated on the ADNI test set. Individual neurologist performance is indicated by the red plus symbol and averaged neurologist performance along with the error bars is indicated by the green plus symbol on both the sensitivity-specificity and precision-recall curves on the ADNI test data. Visual description of pairwise Cohen's kappa (κ), which denotes the inter-operator agreement between all the 11 neurologists is also shown. (B) Sensitivity-specificity and PR curves calculated on the AIBL, FHS and NACC datasets, respectively. 46

Figure 4.6. Visualization of data. (A) Voxel-level MRI intensity values from all four datasets (ADNI, AIBL, FHS and NACC) were used as inputs for a t-SNE plot. The color in the plot represents the site and the marker represents the label. (B) This t-SNE plot was generated only on using the ADNI dataset, where the color was used to represent the scanner. (C) FCN-based outputs that served as input features to the MLP model were embedded in a two-dimensional plot generated using t-SNE for the two classes (AD and NC). The color (blue versus red) was used to distinguish normal cognition from AD cases, whereas a unique symbol shape was used to represent individuals derived from the same cohort. 48

Figure 5.1. Study design of a multi-task deep learning framework. Multimodal data including MRI scans, demographics, medical history, functional assessments, and neuropsychological test results were used to develop deep learning models on various classification tasks. Eight independent datasets were used for this study, including NACC, ADNI, AIBL, FHS, LBDSU, NIFD, OASIS, and PPMI. We selected the NACC dataset to develop three separate models: (i) an MRI-only CNN model (ii) non-imaging models in the form of traditional machine learning classifiers, which did not use any MRI data (iii) a fusion model combining imaging and non-imaging data within a hybrid architecture joining a CNN to a CatBoost model. First, T1-weighted MRI scans were input to a CNN to calculate a continuous DEmentia MOdel (DEMO) score to assess cognitive status on a 0 to 2 scale. For individuals with DE diagnosis, the multi-task CNN model simultaneously discriminated their risk of having AD versus nADD, a classification that we refer to as the ADD task. We denoted the probability of AD diagnosis as the ALZheimer (ALZ) score. 52

Figure 5.2. Data selection. Data from eight distinct study cohorts contributed towards the model development, validation, and testing: The NACC dataset (n=4,822), the OASIS-3 dataset (n=666), the ADNI dataset (n=1,821), the NIFD dataset (n=253), the PPMI dataset (n=198), the AIBL dataset (n=661), the FHS dataset (n=313), and in-house data from the LBDSU (n=182). In each dataset, T1-weighted, 1.5 and 3 Tesla MRIs were selected from participants. Only MRIs gathered within 6 months of MCI, AD or non-ADD diagnosis or last confirmed clinical visit (in the case of NC participants) were included for analysis. 59

Figure 5.3. MRI preprocessing pipeline. MRI scans from all datasets were preprocessed using a common pipeline implemented in FSL. Raw MRIs were first reoriented to a standard axis layout and then aligned to the MNI-152 template using a linear registration tool and automatically identified region-of-interest. These aligned MRIs were then skull-stripped, and the resultant brains then underwent a second linear registration for fine-tuning of MNI alignments, as well as bias field correction for magnetic field inhomogeneities. Finally, specific brain regions were segmented by aligning the Hammersmith Adult brain atlas to registered brains using a non-linear registration. All processed MRIs were inspected visually, and individual brain extraction parameters were adjusted as needed for cases with failed registration. All FSL commands for the above steps are listed within boxes in the accompanying figure. 60

Figure 5.4. Site- and scanner-specific observations. (a) t-distributed stochastic neighbor embedding (tSNE) embeddings of down sampled MRI scans are shown. The down sampling was performed on the post-processed MRI scans using spline interpolation with a down sampling factor of 8 on each axis. (b) tSNEs of hidden-layer activations from the penultimate CNN hidden layer. Individual points correspond to internal representations of MRI scans during testing and are colored by cohort label. (c) tSNE of down sampled MRI scans from the NACC dataset is shown. Individual points are colored by the unique identifier of the Alzheimer Disease Research Centers (ADRCs) that participate in the NACC collaboration. (d) tSNE for penultimate-layer activations colored by ADRC ID are shown. (e) tSNE of down sampled MRI scans from the NACC dataset is shown. Color is assigned based on the manufacturer of the MRI scanner. (f) tSNE of penultimate layer activations is shown for cases in the NACC dataset. Embeddings are equivalent to those visualized in (d) but are now colored by the manufacturer of the scanner. (g) A tabular representation of disease category counts by manufacturer is presented. Only cases from the NACC dataset are included. We provide the Mutual Information Score (MIS) to quantify the correlation between disease type and scanner manufacturer. (h) We also provided a tabular representation of disease category counts stratified by ADRC ID in the NACC dataset. MIS is once again shown to quantify the degree of correlation between diagnostic labels and individual centers participating in the NACC study 62

Figure 5.5. Performance of the deep learning models. (a-b) ROC curves showing true positive rate versus false positive rate and PR curves showing the positive predictive value versus sensitivity on the (a) NACC test set and (b) OASIS dataset. The first row in (a) and (b) denotes the performance of the MRI-only model, the non-imaging model, and the fusion model (CNN and CatBoost) trained to classify cases with NC from those without NC (COG_{NC} task). The second row shows ROC and PR curves of the MRI-only model, the non-imaging model, and the fusion model for the COG_{DE} task aimed at distinguishing cases with DE from those who do not have DE. The third row illustrates performance of the MRI-only model, the non-imaging model, and the fusion model focused on discriminating AD from nADD. For each curve, mean AUC was computed. In each plot, the mean ROC/PR curve and standard deviation are shown as bolded lines and shaded regions, respectively. The dotted lines in each plot indicate the classifier with the random performance level. 68

Figure 5.6. Neuroimaging signatures of dementia. (a-b) SHAP value-based illustration of brain regions that are most associated with the outcomes. The first columns in both (a) and (b) show a template MRI oriented in axial, coronal, and sagittal planes. In (a), the second, third and fourth columns show SHAP values from the input features of the second convolutional block of the CNN averaged across all NACC test subjects with NC, MCI, and dementia, respectively. In (b), the second and third columns show SHAP values averaged across all NACC test subjects with AD and nADD, respectively. (c) Brain region-specific SHAP values for both AD and nADD cases obtained from the NACC testing data are shown. The violin plots are organized per lobe and in decreasing order of mean absolute SHAP values........... 71

Figure 5.7. SHAP-based disease signature of stratified non-AD dementias. We presented the distribution of the regionally averaged SHAP values from the subjects that were correctly predicted as AD or non-AD dementias. The x-axis contains the region names that each violin plot is corresponding to. Comparisons on the regionally averaged SHAP distribution were made between AD and each non-AD dementias, including frontotemporal dementia (top row), Lewy body dementia (middle row), and vascular dementia (bottom row).. 73

Figure 5.8. Network analysis of saliency maps over brain regions. The networks were generated using all (a) AD and (b) nADD subjects, respectively. We selected 33 representative brain regions for graph analysis and visualization of sagittal regions, as well as 57 regions for axial analyses. Nodes representing brain regions are overlaid on a two-dimensional brain template and sized according to weighted degree. The color of the segments connecting different nodes indicates the sign of correlation and the thickness of the segments indicates the magnitude of the correlation. It must be noted that not all nodes can be seen either from the sagittal or the axial planes.. 75

Figure 5.9. Feature importance. (a-b) Fifteen features with highest mean absolute SHAP values from the fusion model are shown for the COG and ADD tasks, respectively across cross-validation rounds (n=5). Error bars overlaid on bar plots are centered at the mean of the data and extend +/- one standard deviation. For each task, the MRI scans, demographic information, medical history, functional assessments, and neuropsychological test results were used as inputs to the deep learning model. The left plots in (a) and (b) illustrate the distribution of SHAP values and the right plots show the mean absolute SHAP values. All the plots in (a) and (b) are organized in decreasing order of mean absolute SHAP values. (c-d) For comparison, we also constructed traditional machine learning models to predict cognitive status and AD status using the same set of features used for the deep learning model, and the results are presented in (c) and (d), respectively. The heat maps show fifteen features with the highest mean absolute SHAP values obtained for each model.......................... 78

Figure 5.10. Expert-level validation. (a) For the $\mathrm{COG}_{\mathrm{NC}}$ task (first row), the diagnostic accuracy of board-certified neurologists (n=17) is compared to the performance of our deep learning model using a random subset of cases from the NACC dataset (n=100). Metrics from individual clinicians are plotted in relation to the ROC and PR curves from the trained model. Individual clinician performance is indicated by the blue plus symbol and averaged clinician performance along with error bars is indicated by the green plus symbol on both the ROC and PR curves. The mean ROC/PR curve and the standard deviation are shown as the bold line and shaded region, respectively. A heatmap of pairwise Cohen's kappa statistic is also displayed to demonstrate inter-rater agreement across the clinician cohort. (a) For the $\mathrm{COG}_{\mathrm{DE}}$ task (second row), ROC, PR, and interrater agreement graphics are illustrated with comparison to board-certified neurologists in identical fashion to (a). For both (a) and (b), all neurologists were granted access to multimodal patient data, including MRIs, demographics, medical history, functional assessments, and neuropsychological testing. The same data was used as input to train the deep learning model. 79

Figure 5.11. Validation against expert assessments on atrophy. SHAP values from the second convolutional layer averaged from selected brain regions are shown plotted against atrophy scores assigned by neuroradiologists. Orange and blue points (and along with regression lines and 95% confidence intervals) represent left and right hemispheres, respectively. Spearman correlation coefficients and corresponding two-tailed p-values are also shown and demonstrate a statistically significant proportionality between SHAP scores, and the severity of regional atrophy assigned by clinicians. 81

Figure 5.12. Neuropathological validation. (a) An example case from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset is displayed in sagittal, axial, and coronal views. The SHAP values derived from the second convolutional block and neuropathologic ABC scores are mapped to brain regions where they were measured at the time of autopsy. Visually, high concordance is observed between anatomically mapped SHAP values regardless of the hidden layer from which they are derived. Concordance is observed between the SHAP values and neurofibrillary tangles (NFT) scores within the temporal lobe. (b) A heatmap is shown demonstrating Spearman correlations between population-averaged SHAP values from the input features of the second convolutional layer and stain-specific ABC scores at various regions of the brain. A strong positive correlation is observed between the SHAP values and neuropathologic changes within several areas well-known to be affected in AD such as the hippocampus or parahippocampus, amygdala and temporal gyrus. (c) Beeswarm plots with overlying box-and-whisker diagrams are shown to denote the distribution of ABC system sub-scores (horizontal axis) versus model-predicted cognitive scores (vertical axis). The displayed data points represent a pooled set of participants from ADNI, NACC, and FHS (n=118) for whom neuropathology reports were available from autopsy. Each symbol represents a study participant, boxes are centered at the median and extend over the interquartile range (IQR), while bottom and top whiskers represent 1st and 3rd quartiles -/+ 1.5 x IQR, respectively. Note: We denote $p < 0.05$ as *; $p<0.001$ as **, and $p < 0.0001$ as *** based on post-hoc Tukey testing. (d) A heatmap demonstrating the distribution of neuropathology scores versus model predicted AD probabilities. Herein, each column within the map represents a unique individual whose position along the horizontal axis is a descending function of AD risk according to the deep learning model. The overlying hatching pattern represents the dataset (ADNI, NACC and FHS), from which everyone is drawn. 83

Figure B.1. Non-imaging features missing rate. The proportion of missing data is shown for all non-imaging features across the eight cohorts. A value of 0.0 represents that no data is missing, while a value of 1.0 indicates that all data for a particular feature was absent. We further stratify missingness by diagnostic label (NC, MCI, AD, and nADD) to demonstrate instances in which data-availability and disease status may be correlated. 104

Figure C.1. MRI model performance on AD task across nADD sub-groups. The ROC and PR curves demonstrate the ability of the MRI-only model to accurately delineate AD from nADD cases when evaluated across several nADD subgroups, including vascular dementia (VD), Lewy body dementia (LBD), and frontotemporal dementia (FTD). The area under ROC and PR curves demonstrate that model performance is the strongest on non-Parkinsonian dementias (VD and FTD) than Parkinsonian dementias (PDD and LBD). Performance from the (a) NACC test set and (b) all external datasets are shown. 105

Figure C.2. Fusion model performance on AD task across nADD sub-groups. The ROC and PR curves demonstrate the ability of the fusion (CNN and CatBoost) model to accurately delineate AD from nADD cases when evaluated across the nADD subgroups. The area under ROC and PR curves demonstrate that the inclusion of non-imaging data elevates the model's performance on Parkinsonian dementias. Performance from the (a) NACC test set and (b) OASIS dataset is shown. 106

Figure D.1. MRI model's performance on all external data. ROC curves (row 1) on $\mathrm{COG}_{\mathrm{NC}}$ task (left), $\mathrm{COG}_{\mathrm{DE}}$ task (middle) and ADD task (right). PR curves (row 2) on $\mathrm{COG}_{\mathrm{NC}}$ task (left), $\mathrm{COG}_{\mathrm{DE}}$ task (middle) and ADD task (right). 107

Figure F.1. Non-imaging and fusion models with partial features. We assessed the performance of 4-way classification on non-imaging and fusion (CNN + CatBoost) models when using varying combinations of historical ("his"), neuropsychological ("np"), and functional ("func") variables (as well as MRI-derived variables in the case of the fusion model). The panel (a) on the left shows the models' performance using different feature combinations but without MRI information. The panel (b) on the right shows the model performance using various feature combinations with MRI included. The model accuracy, F-1, sensitivity, specificity, and MCC values are demonstrated, and comparison is made between the NACC test set and the OASIS dataset for each performance metric. Of note, similar distributions of performance metrics are observed between the two datasets, thus suggesting that the model does not privilege particular features in one dataset over the other. 111

List of Abbreviations

AD	Alzheimer’s disease
ADNI	Alzheimer’s Disease Neuroimaging Initiative
ADRC	Alzheimer’s Disease Research Center
AIBL	Australian Imaging, Biomarker and Lifestyle Flagship Study of Aging
ANN	Artificial neural network
ANOVA	Analysis of variance
APOE	Apolipoprotein E
CAM	Class activation mapping
CNN	Convolutional neural network
DE	Dementia
DL	Deep learning
DPM	Disease probability map
FCN	Fully convolutional network
FDA	Food and Drug Administration
FHS	Framingham Heart Study
GP	General practitioner
KNN	K-nearest neighbor
LBD	Lewy Body Dementia Center for Excellence at Stanford University
LIME	Local interpretable model-agnostic explanations
MAE	Mean absolute error
MCC	Matthew correlation coefficient

MCI	Mild cognitive impairment
ML	Machine learning
MLP	Multi-layer perceptron
MMSE	Mini-mental state examination
MRI	Magnetic resonance imaging
MSE	Mean square error
NACC	National Alzheimer’s Coordinating Center
nADD	non-Alzheimer’s disease dementia
NC	Normal cognition
NIFD	Frontotemporal Lobar Degeneration Neuroimaging Initiative
OASIS	Open Access Series of Imaging Studies
PET	Positron emission tomography
PPMI	Parkinson’s Progression Marker Initiative
PR	Precision-recall
rMSE	root mean square error
RNN	Recurrent neural network
ROC	Receiver operating characteristic
SHAP	SHapley Additive exPlanations
tSNE	t-distributed stochastic neighbor embedding
UDS	Uniform Data Set
VD	Vascular dementia

Chapter 1

Introduction

1.1 Alzheimer’s Disease and Dementia

Dementia is a global epidemic and itself is an overall term for a collection of symptoms including declined memory and impaired language, execution, and reasoning abilities[1]. There are various brain diseases and disorders that can cause dementia. Alzheimer’s disease (AD) is the leading cause of dementia, which accounts for 60-70% of cases[2]. Other common causes of dementia are Lewy Body dementia, vascular dementia, frontotemporal dementia, Parkinson’s disease etc. Multiple etiologies of dementia may occur simultaneously within the brain which contribute to a condition called mixed dementia[3]. The tangled landscape of dementia complicates diagnostic and treatment procedures, which consequently limits the effectiveness of containing disease prevalence and the already burdensome morbidity and mortality rate among the elderly group[4].

AD is a chronic progressive neurodegenerative disease that is associated with certain brain pathological changes. Individuals with early symptoms may occasionally encounter difficulties conducting some daily tasks. As the disease progresses, the family of the subjects suffering from worsening cognitive impairment may start to notice typical signs of AD from the subjects, featuring declined memory, shortened attention span, language difficulties etc. At a late stage of disease progression, individuals with AD may not be able to recognize family members and may get severely injured due to the loss of balance and motion. The rate of progression for AD varies from several years to more than a decade before it becomes fatal. Such dire consequences of AD, especially at the late stage, makes it the sixth leading cause of death in the United States[5].

The underlying mechanisms driving AD progression remains an open research problem. Researchers have demonstrated that the accumulation of certain abnormal proteins, i.e., beta-amyloid plaques and tau tangles, inside the brain can cause neuronal degeneration, and these misfolded proteins have been considered as the hallmark pathologies of AD[6]. Aging is also a risk factor of AD[7]. Older people tend to have a larger degree of brain volume loss and are subject to a higher risk of having impaired cognition[8]. The subtle structural changes in the brain due to normal aging or AD makes diagnosis a challenging task even for expert clinicians familiar with reviewing various forms of data including medical scans like magnetic resonance imaging (MRI). Some genetic characteristics, like the presence of polymorphism in the apolipoprotein E (APOE), have also been considered as a risk factor for AD[9].

About 11% of adults of age 65 or older suffer from AD in the United States[10]. As the population ages, the number of people with AD is projected to nearly triple from 2010 through 2050[11]. The disease prevalence in combination with the chronic nature of the disease makes AD a burdensome healthcare challenge despite heavily invested economic resources and community support for dementia care. This projection accentuates the need to find better diagnosis, treatment, and management strategies for AD and alerts us to prepare for the worsening future shortage of dementia specialists. The current dementia workforce consists of different types of medical professionals, including well-trained specialists like neurologists and neuroradiologists, and primary care physicians within communities[12]. Many general practicing physicians reported that they were not confident on some of the diagnoses they made given limited patient information. Sometimes, patients may be referred to a dementia specialist to confirm the diagnosis using more comprehensive clinical assessments, including neuropsychological tests, functional questionnaires, depression evaluation, medical imaging scans etc. However, due to the sparsity of dementia specialists, most of the individuals first diagnosed with dementia are diagnosed by a primary care physician, and there are many people with dementia who remain undiagnosed in their life. If AI-aided models installed in an upstream clinical setting (i.e., primary care facility) can be demonstrated to diagnose at the same level as that of the experts in memory clinics, these automatically operating diagnostic models can make a broad and profound impact on AD management provided them being accepted by the healthcare community.

Tremendous efforts have been made to develop medications for AD. There are several drugs that have been approved by the U.S. Food and Drug Administration (FDA)[13]. Some of them intend to slow down the progression of AD symptoms, while the recently approved Aducanumab addresses the underlying biological mechanism of the disease by reducing beta-amyloid plaques from the brain[14]. No matter which type of treatment is used, the effectiveness of the treatment can be dramatically decreased if not being applied at the right time[15]. The market demand for timely diagnosis in combination with the dearth of experts makes the scalable AI-aided diagnostic solution favorably and urgently needed to contain the prevalence of AD and dementia.

1.2 Machine Learning and Deep Learning

Machine learning (ML) is a sub-field of AI where computer algorithms are capable of automatically obtaining knowledge from data and solving a variety of tasks without being explicit programmed as conventional algorithms work. The root of ML is statistics which is a broader term for the principle and methodology of analyzing, interpreting, and understanding patterns of data. Tom Mitchell provided a formal definition of ML algorithms in his book[16]: "Computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.". According to this definition, learning happens with experiencing real-world problems and getting constant feedbacks on distinct responses. The real-world experience which is reflected by data needs to be collected, organized, and annotated in a way that can be digested by a computer and follows machine learning paradigms.

To improve the performance of ML models, one fruitful direction is to improve the quality and quantity of data. The more sophisticated model you have, the more data you need to feed into the model to acquire a decent performance. The rapidly increased number of digitized records provides an unprecedented opportunity for AI to reshape many aspects of the society, e.g., healthcare, education, finance, marketing etc. The quality of data also matters since learning from improperly labeled or mistakenly sampled data can often jeopardize a model's generalizability, fairness, and robustness.

One of the fundamental challenges of ML research is to design and develop new learning algorithms under the theoretical guidance of statistical laws to improve the capacity and efficiency of the learning. It is almost impossible for a programmer to provide step-by-step instructions to a model to solve nontrivial tasks especially when dealing with complex data like images, voice, video, text etc. Instead, provided that high-level instructions like the learning goal and the optimization method are specified by a user, a machine learning model can automatically learn by experiencing the data. I will introduce some of the mainstream optimization algorithms in the following chapter, including gradient descent[17], genetic algorithm[18], and simulated annealing[19], which all incrementally update a model's parameters to improve the performance as quantified by the metrics.

Based on the problem formulation, there are different ML paradigms, i.e., supervised learning, unsupervised learning, reinforcement learning, and other less common types[20]. Commonly used ML methods are under the category of supervised learning where the data is presented as a collection of input x and label y pairs. A supervised model learns the mapping f from input feature x to the ground truth label $y = f(x)$ and is awarded or penalized depending on whether the prediction corresponds to the label. There could be different types of labels, including discrete labels for classification or continuous labels for regression tasks.

Traditional ML models, e.g., support vector machine, decision tree, random forest, and nearest neighbor, have been commonly applied to process well-organized tabulate information. However, it has been demonstrated that a variety of deep learning (DL)[21] models are more suitable and capable of handling complicated data like image, text, video, and audio. The simplest form of a deep neural network is a multi-layer perceptron (MLP)[22] which is also known as the fully connected network. Convolutional neural network[23]

(CNN) is another type of deep neural network that has been broadly trained to tackle vision-related tasks like image classification, interpretation and object detection. Recurrent neural networks[24] (RNN) are more capable of handling sequencing data for tasks like time series forecasting, language understanding and translation.

Over the past decade, we have seen many successful industrial ML applications including personalized content recommendation, AI-powered search engines, AI-driven text autocompletion and grammar autocorrection, fraud detection for transactions, and many other use cases. Besides the commercial success of AI applications, ML also sheds new light on the research areas of foundational science, for instance, using ML to predict protein structures[25] and gene expression[26], and innovating density functional computation with DL[27]. Though ML has been progressed expeditiously, there still remain concerns rooted from the lack of trust, transparency, fairness, accountability, and other ethical aspects of leveraging ML model predictions[28]. Society has also been paying attention to setting regulatory hurdles for ML applications and staying cautious while accepting ML as part of technological advancements.

1.3 Interpretable Machine Learning

Though ML models have been demonstrated with strong predictive power in many application domains, most of them do not explain how the predictions were made which consequently prevents them from earning the trust from domain experts and the community. In addition to merely leveraging highly accurate "black box" predictions, we have recently witnessed an increased interest and attention on interpreting a model's predictions[29,30]. The concept of interpretable ML can be confusing due to the lack of well-formed definition and the existence of a broad variety of interpretable ML methods[31–36]. Generally speaking, a model is interpretable if it can provide relevant domain knowledge or information regarding its predictions[30]. For example, in the application of medical image analyses, it would be more useful if a model can identify the regions from a medical scans that are indicative of certain disease[37]. The ability of interpreting a model's predictions also greatly benefits the development of ML models by preventing apparent mistakes using clues from interpretable outcomes. In addition, when ML is used for scientific research, interpretability also plays a critical role in extracting new patterns from data and confirming existing knowledge with data-driven findings.

Interpreting a linear regression model with tens of features is an easy task, considering the weights from the linear model directly indicate the importance of features. As the complexity of the model increases, the causality between input features to final predictions becomes intractable for a human. Going beyond simple statistical models, many ML models incorporate non-linear operations into complicated model architectures which boost predictive accuracy at the cost of losing interpretability. There are two well-established routes that introduce interpretability to an ML model[30]. The first one is through designing and developing intrinsically interpretable models beforehand, for example, rule-based ML models, decision trees, K-nearest neighbor (KNN), and Naïve Bayes model. Alternatively, we can analyze the interpretability as a post hoc step on an already trained model[31–33,36]. Post hoc methods are often model agnostic, since the development phase is decoupled from the interpretation phase, and thus allows better flexibility in model selection.

Chapter 2

Machine Learning and Deep Learning

2.1 Traditional Machine Learning Models

Before the prevalence of deep neural networks, researchers developed many classical machine learning algorithms that are still being frequently used, including Markov chain[38], nearest neighbor[39], support vector machine[40], decision tree[41], random forest[42] and etc. These classical algorithms are usually more data- and computation-efficient, and easier to interpret compared with deep neural networks. Though we have seen the trend that the predictive power of classical algorithms gets surpassed by deep neural networks, the idea behind those classical algorithms still profoundly benefits the development of state-of-the-art deep learning models and industrial machine learning frameworks. For instance, the approximate nearest neighbor algorithm[43] has been widely used as one of the core component of industrial recommendation systems. Gradient boosting tree[44,45] is another successful example of integrating the classical tree-based models with the gradient descent optimization method ubiquitously used in deep learning.

2.2 Artificial Neural Network

Artificial neural network (ANN) is a subset of ML models[46] whose development was inspired by the mechanism of neuronal connections in human brain. The simplest form of ANN, i.e., the perceptron, was invented at the Cornell Aeronautical Laboratory in 1958[47].

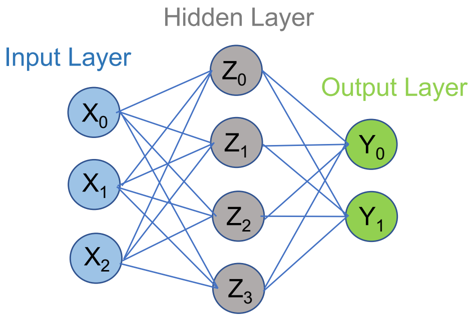


Figure 2.1. Structure of a perceptron model.

A perceptron model comprises an input layer, one or more hidden layers and one output layer. Each connection between two nodes is associated with a weight, and the value of an intermediate node can be derived using the formula below:

$$Z_i = ReLu\left(\sum_{j=0}^{n} X_j W_{ij} + b_i\right) \qquad (1)$$

where W_{ij} is the weight and b_i is the bias for a node Z_i. $ReLu$ is one type of non-linear activation functions which sets negative input value to zero and keeps the positive input value the same[48]. Alternatively, other non-linear activation functions have also been commonly used, e.g., sigmoid, hyperbolic tangent etc. Without non-linear activations, a perceptron model can only represent a linear function no matter how many linear layers are stacked together. Several works[49,50] proofed that multi-layer perceptron is a universal approximator with the capability of approximating a broad type of functions, provided that proper weights were given. In the following sections, other types of ANNs are introduced.

2.3 Convolutional Neural Network

Convolutional neural network (CNN) is another type of ANN which has been commonly used for vision-related tasks like image classification, interpretation, segmentation, object detection etc., while it can also be used to analyze data of other modalities, e.g., text, voice, or other sequencing data. The core component of a CNN is the convolutional layer which contains multiple kernels - small matrices with user defined shape - as its parameters. The convolution operation is typically conducted between an input feature and a kernel. To be more specific, the operation slides a kernel through the input feature, and at each location, the summation of the element-wise multiplications between the input patch and the kernel is stored as the output. Mathematically, this operation is equivalent to the inner product between the image patch and the kernel. Thus, the output values of the convolution operation reflect the similarity between input patches and the kernel. To capture distinct spatial patterns from the image, a convolutional layer usually has tens or hundreds of kernels to guarantee a sufficient learning capacity.

The major advantage of using convolution instead of fully connected layer is that a convolutional layer only contributes a small number of parameters to the model, and thus makes that model less prone to over-fit the data. If we apply multi-layer perceptron on an image with N-by-N pixels, the number of parameters required within a layer is in the order of $O(N^2M^2)$ assuming the output feature map is of shape M-by-M. In contrast, the number of parameters within a convolutional layer equals $n_k k^2$, where n_k is the number of kernels and k^2 is the size of a kernel. Thus, the convolutional layer provides an efficient way of sharing weights/parameters over all spatial locations which, therefore, regularizes the complexity of a neural network[51].

Another commonly used operation within a CNN is called pooling which is defined as the process of scanning a window throughout the input feature and outputting the max or the average feature value within the window. Pooling can be used to effectively decrease the size of features, while still passing salient features to the successive layer. Researchers have made significant progress towards exploring new CNN architectures for a broad range of computer vision tasks[52–56]. The figure below shows the structure of the VGG model which was developed by the Visual Geometry Group at Oxford University[52] and has been widely used as a baseline model to compare with. There is an ongoing trend of using deeper CNN models with up to hundreds of convolutional layers, for example the ResNet-152[56], to fully leverage the predictive power of deep neural networks.

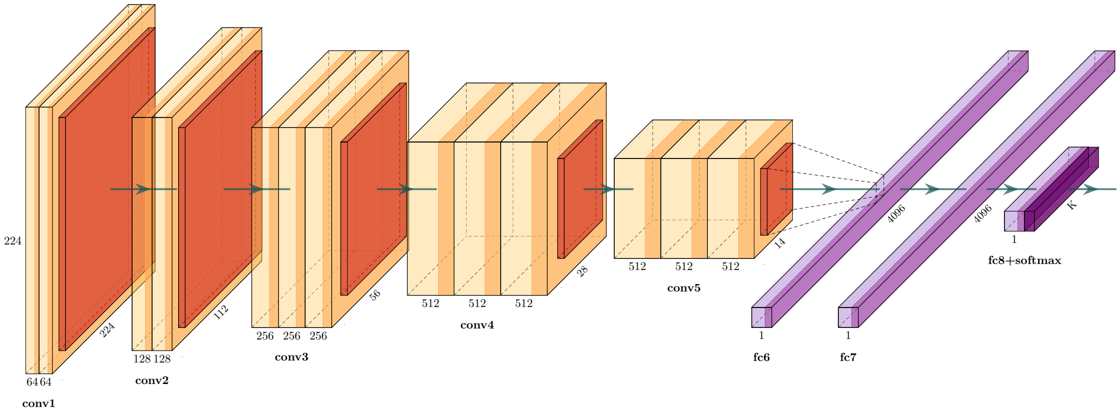


Figure 2.2. Architecture of a convolutional neural network.

For a CNN that is used for classification, one or more fully connected layers are usually appended after the convolutional layers to convert the feature map from the last convolutional layer to predictions. Researchers also proposed fully convolutional network (FCN) for the task of semantic segmentation[57] which only comprises convolutional layers. One advantage of the FCN is that the model can accept input with arbitrary shape. And the size of the output depends on the input size. Since both convolutional layer and the fully connected layer are linear functions, it is worth mentioning that a fully connected layer has an equivalent convolutional version. For example, the convolutional layer with kernel size 1 by 1 is identical to a fully connected layer whose input and output dimension equal to the input and output channel of the convolutional layer, respectively. Similarly, the fully connected layer that directly follows a convolutional layer is also equivalent to a convolutional layer whose kernel size is identical to the spatial dimension of the preceding feature map. Thus, a CNN classifier with fully connected layers can be converted to an equivalent FCN. The parameters of a CNN are not predetermined by humans to detect certain edges or corners, instead these parameters are being updated automatically according to the rule of learning which will be discussed in the next section.

2.4 Optimization and Gradient Descent

Training a neural network is an optimization problem wherein the best solution is selected from a broad set of candidates based on a performance criterion. In the context of ML, a cost function usually corresponds closely with model performance, thus allowing the optimization process to be guided towards the direction of descending value of the cost function. We only discuss the minimization of the cost function here, and we will use the term loss function interchangeably with cost function. For a regression problem like predicting a house price, commonly used cost functions are mean square error (MSE), mean absolute error (MAE), Huber loss[58], and log hyperbolic cosine loss. For classification, the most often used loss function is cross-entropy loss[59]. Qi Wang et al provided a comprehensive survey of loss functions for training machine learning models[60].

To solve the optimization problem, researchers have developed many numerical methods including convergent iterative methods such as Newton's method, gradient descent, and some heuristic methods like genetic algorithm[61] and simulated annealing[19,62]. Partly due to the computational burden of calculating Hessian and the inadequacy of escaping from proliferated saddle points[63], the second-order optimization methods such as Newton's method have not been used as common as the first-order gradient-based approaches. Nonetheless, gradient descent method has its own limitations, i.e., the slow convergence when gradient diminishes at local minimum and the susceptibility to exploding gradients.

Mathematically, the weights w of a model can be optimized using the formula below:

$$w_{i+1} := w_i - \eta \nabla_w l(w_i, x) \tag{2}$$

where the subscript of w represents the iteration step, η is the learning rate and $l(w_i, x)$ is the loss function given weights w and input x. To avoid being trapped in local minimums, the stochastic version of gradient descent was developed which uses the gradient estimated from a few samples to train the model. From the past decades, we witnessed an increased efficiency in training a neural network using the stochastic gradient descent method. Also, recent progress has been made towards incrementally improving the gradient descent optimizers[64]. One popular variant of the optimizer uses accumulated momentum of gradients from previous iterations to accelerate model convergence[65]. Another variant which has been widely used in the optimizers like Adam[65], RMSprop, and Adagrad[66], adaptively adjusts the learning rate so that the training process can stay efficient no matter whether the gradient landscape is steep or flat.

2.5 Model Training and Evaluation

2.5.1 Splitting the Data

To train and evaluate a ML model, we need to firstly split the data into non-overlapping segments: the training set, the validation set, and the testing set. The model will be trained on all instances from the training set using one of the optimization methods described above. The validation set is commonly used to monitor the model's performance during the training process so that the model's weights with the best performance on the validation set can be stored as the final model. The testing set is the ideal segment to be used to report the model's performance as none of the testing instances have been used for model training and selection. It is worth noting that "data leakage" between training, validation, and testing sets can occur due to lack of observation or other human errors. If leakage happened, the observed performance on the testing set may be superficially high since many cases have already been seen during the training stage.

2.5.2 Overfitting and Regularization

There are 2 major sources of error for machine learning models: bias and variance[67]. A model that is too simple may underfit the data and predict with large bias. On the contrary, an over complicated model that can easily overfit the training set may perform poorly on the testing set due to the high variance of the prediction. In the deep learning era, overfitting is becoming a widely existing problem due to the trend of increasing model complexity. To prevent such a problem, researchers have developed several regularization techniques to limit the complexity of the model. For example, Lasso and Ridge regression are two types of linear methods that are regularized with different forms of penalty on model's complexity.

Dropout[68] is an operation that randomly sets node values to zero according to a specified probability during the training stage. Due to the randomness of dropping out nodes, a single model can be used to simulate many distinct models, each of which has a different combination of dropped-out nodes. Since a portion of nodes are set to zero, the model becomes less complicated, and the model size is proportional to the dropout probability. The behavior of the dropout operation is different during the testing stage. Instead of relying on any randomly dropped-out model, it is a better practice to average the predictions from all randomly dropped-out models. To achieve that, one just needs to multiply the dropout probability to all node values. In summary, the dropout operation not only regularizes the model complexity but also harnesses the predictive power from a big ensemble of random models.

BatchNorm[69] is another widely adopted technique that improves the training of neural networks by rescaling the distribution of the intermediate features. Though the effectiveness of BatchNorm is well-established, the underlying reasons for why BatchNorm improves the training of neural networks remain debatable. The original paper claimed that BatchNorm effectively reduces the internal covariate shift of the feature values, thus allowing larger learning rate and less-stringent weights initialization[69]. Other explanations for this aspect also exist. Santurkar et al. attribute the effectiveness to the smoother optimization landscape induced by the rescaling effect of BatchNorm[70].

2.5.3 Evaluation

In this section, we provide a comprehensive list of performance metrics along with approaches of estimating the variance/confidence interval of these metric values. For a binary classification task, we can characterize a model's performance using the following metrics: accuracy, precision, recall, sensitivity, specificity, F-1 score, Matthew correlation coefficient (MCC), receiver operating characteristic (ROC) curves, and precision-recall (PR) curves. For a multi-way classification, the prediction task can be decoupled into many binary classifications by considering one class as a group and all remaining classes as the other group. Thus, one can report the metric values for each of the class-specific binary classifications and the overall metrics averaged over all class-specific metrics. For regression tasks, commonly reported metrics include mean square error (MSE), and root mean square error (rMSE).

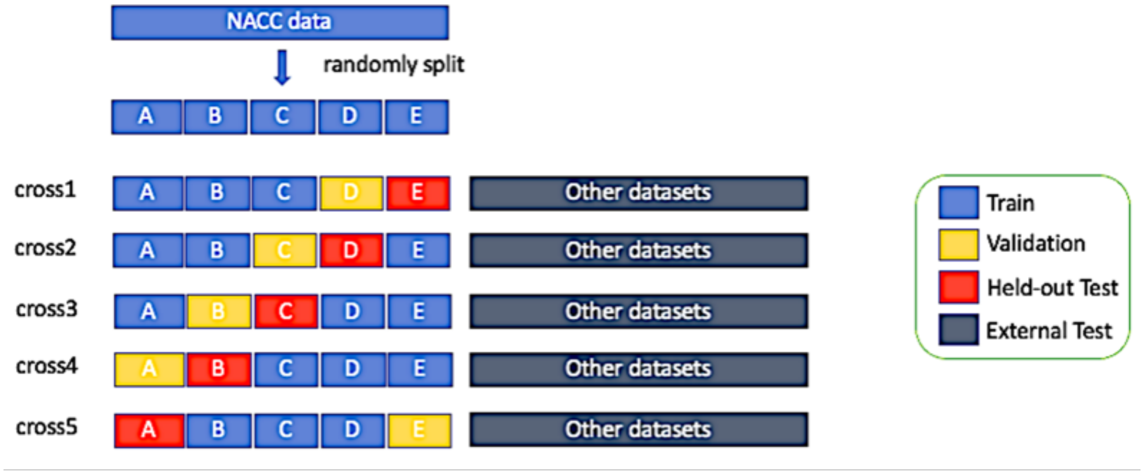


Figure 2.3. Splitting data for cross-validation.

To properly report the variance of model performance, it is necessary to run ML experiments multiple times using strategies like boot-strapping or cross-validation. To conduct cross-validation, one can first split the randomly shuffled data into *k* groups where one group can be used to hold-out testing and remains groups can be used for training and validation. By rotating the assignment of groups on training, validation, and testing, one can conduct *k* independent experiments on which variance of model performance can be estimated. With this rolling scheme (Figure 2.3), every group has a chance of being used for testing.

Chapter 3

Interpretable Machine Learning

3.1 Introduction

Machine learning has made substantial progress across different application domains. As the popularity of ML increases, the black-box nature of ML models becomes one of the major obstacles to use ML predictions in high-stakes scenarios. Before reaching the stage where ML models can be used as a clinical tool, it is crucial to rigorously validate the model and obtain a deep understanding of the model's decision-making process. It is an often-seen argument that explainable AI, which elucidates a model's decision-making process and unveils unwanted prediction biases, can build trust with healthcare workers, and increase their willingness to use AI-aided systems. However, some researchers believe this argument brings a false hope for explainable AI[71], as current methods are unable to deliver high-fidelity patient-specific explanations. It is still debatable whether an AI-aided model should have explainability as one of the requirements for deployment, and it is not clear whether doctors will be biased from over-trusting and misinterpreting some of the explainable results.

Though it is almost impossible at this stage to explain exactly how a sophisticated ML model makes its predictions, it is feasible to estimate the causal relationship between input features and model predictions. Interpretable ML is a research field where researchers focus on extracting meaningful knowledge or insights on the dependency between model predictions and feature values. With the capability of interpreting a model, researchers or engineers who develop the model can better understand the model, thus allowing them to debug, verify, and refine the system. From the perspective of consumers who use ML predictions to achieve their own goal, a better decision can be made based on a more holistic view of the ML insights, beyond just model predictions, from interpretability analyses. Thus, it is still an important research direction to pursue more accurate and easier-to-understand ways of interpreting ML predictions. In this chapter, I will introduce a broad range of interpretable ML methods and discuss both the advantages and limitations of these approaches.

Interpretable ML methods can be categorized in multiple ways. Firstly, based on how the interpretation is generated, the methods can be divided into two categories: intrinsic interpretable methods and post hoc interpretable approaches. Intrinsically interpretable ML models, like decision trees, linear models, and Bayesian models, are themselves interpretable as the information flow within such models is easy to follow. However, for a more complicated model that is not intrinsically interpretable, post hoc methods, like LIME[36] and SHAP[31], can be applied to derive meaningful insights on relating input and prediction. Secondly, interpretable ML methods can also be categorized according to whether the interpretable outcome is based on a single instance or all instances from a cohort. Local interpretable ML methods provide instance-specific insights, whereas global interpretable ML methods, including permutation feature importance[72] and partial dependence analysis[73], only generate interpretable conclusions at the level of the whole dataset. By considering all instance-specific interpretations, local interpretable ML methods can also be used to draw global insights, but it might not be feasible for the opposite direction. Thirdly, interpretable ML methods can also be grouped into model-specific approaches, like class-activation mapping[35], DeepLIFT[74] and integrated gradients[75], and model-agnostic methods such as LIME, SHAP, counterfactual explanations[76].

To guide the selection of interpretable ML methods, Murdoch et al. provided three considerations: predictive accuracy, descriptive accuracy, and relevancy[30]. Predictive accuracy describes how accurate the original ML model is, whereas descriptive accuracy measures how accurate the interpretation is generated. Intrinsically interpretable models usually have low predictive accuracy due to their limited complexity but high descriptive accuracy since the chance of misinterpretation is low given a straightforward workflow. In the case of applying model-agnostic interpretable methods, the predictive accuracy can be high owing to the flexibility of model selection, but the fidelity of the interpretation is often skeptical especially when the interpretation is coming from surrogate models or some post hoc analyses. The relevancy of the interpreted outcomes to the application domain is also important and should be considered while comparing different interpretable approaches.

3.2 Intrinsic and Post hoc Interpretability

3.2.1 Intrinsically Interpretable Models

In this section, I will first introduce a list of intrinsically interpretable ML models and then discuss some post hoc approaches. The simplest interpretable ML model is the linear model as below:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_n x_n \qquad (3)$$

where y represents the target label and x_i and β_i are the input feature and model weight, respectively. The corresponding weight directly implies the impact of the change of a feature value Δx_i. Thus, linear model is a typical example of having low predictive accuracy but high descriptive accuracy. Decision tree is a rule-based model and is presented as a tree data structure whose nodes represent the tests on input features and branches represent the test outcomes (Figure 3.1).

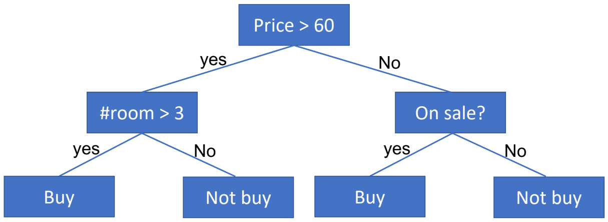


Figure 3.1. A decision tree.

There are a variety of algorithms that can be used to construct a decision tree[77]. The greedy approach is commonly used to grow the tree from root to leaves according to some measurements of information gain like Gini impurity[78] so that the attribute that contains the most information can be firstly included as part of the tree. In combination with other techniques like boosting and bagging, variants of decision trees, such as XGBoost[45] and CatBoost[44], were demonstrated to have much stronger predictive power.

Naïve Bayes is another algorithm that is intrinsically interpretable owing to its oversimplified assumption that all features are independent. A typical classification problem can be formulated as calculating the probability of class C_i given a feature vector $\vec{x} = (x_1, x_2, \dots, x_n)$. According to the Bayes' theorem, the probability can be decomposed as:

$$P(C_i \mid x_1, x_2, \dots, x_n) = \frac{P(C_i)P(x_1, x_2, \dots, x_n|C_i)}{P(x_1, x_2, \dots, x_n)} \tag{4}$$

Under the assumption of feature independence, the numerator of the formula above which is what we are interested to calculate can be simplified to:

$$P(C_i)P(x_1, x_2, \dots, x_n|C_i) = P(C_i)\prod_i P(x_i|C_i) \tag{5}$$

where $P(x_i|C_i)$ can be easily estimated if it is safe to assume that x_i follows certain probability distribution, like Gaussian distribution or Bernoulli distribution.

3.2.2 Post hoc Interpretable Methods

LIME (Local Interpretable Model-agnostic Explanations) is one of the popular post hoc interpretable methods that uses an interpretable surrogate model, like a decision tree or linear model, to approximate the original model around a local prediction. To train such a surrogate model, only the input and prediction pairs are needed from the original model which makes such method universally applicable to any type of models.

A surrogate model g might be able to approximate the original model f with higher fidelity as the complexity of g increases while the interpretability of g decreases. The author of LIME proposed a cost function that considers the fidelity-interpretability trade-off.

$$g = \underset{g}{argmin}\, L(f, g, \pi_x) + \Omega(\mathrm{g}) \tag{6}$$

Where L measures the fidelity of g approximating f in a local domain defined by π_x, and $\Omega(\mathrm{g})$ characterizes the complexity of g. LIME uses instances z that are sampled around the target instance x as data to train the surrogate model with a fidelity function L defined as below:

$$L = \sum_{z} \pi_x(z)(f(z) - g(z))^2 \tag{7}$$

where the π_x is the kernel function that has larger value for nearby perturbed samples.

The Shapley value[79] from cooperative game theory provides a different pathway for producing post hoc local interpretations. The Shapley value is a method for distributing the total profit obtained from a coalition of players to everyone in a fair way which is like attributing a model's prediction to the contribution from each feature. The classical Shapley value for a certain player is defined as the difference between the averaged profit over all coalitions with the player included and that with the player excluded (see the formula below).

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \tag{8}$$

Where ϕ_i is the Shapley value for the *i* th player, $|S|$ is the size of the coalition, and $|F|$ is the total number of players. Though the Shapley value was proofed to have many desired properties, i.e., efficiency, symmetry, linearity and etc.[79], it is extremely computationally expensive to calculate the exact Shapley value due to the exponentially increasing number of feature combinations.

Lundberg et al. proposed an efficient way of estimating the Shapley value by solving a weighted linear regression problem where features were masked out to emulate the absence of players[31]. The authors proofed that with the loss function defined below, the solution of the linear regression yields a good approximation of the Shapley value.

$$L(f, g, \pi_x) = \sum_{z} \pi_x(z)[f(h(z)) - g(z)]^2 \tag{9}$$

$$\pi_x(z) = \frac{M - 1}{\binom{M}{|z|} |z|(M - |z|)} \tag{10}$$

Where *z* is the binary vector to characterize what features are included and the function *h* converts binary vector to the original feature space by replacing ones with the true feature values and zeros with the features' average. To distinguish from the classical Shapley value, we use the term "SHAP" value to represent the outcome of this kernel-based approach. A positive SHAP value indicates a positive contribution towards the final prediction given the current feature value. Since the SHAP value follows the property of additivity, the sum of all the SHAP values over all features should be equal to the final prediction subtracting the baseline value, i.e., cohort-averaged prediction value.

3.3 Global Feature Importance

One of the major outcomes from interpretability analysis is the overall importance of each feature with respect to a ML model. The ranking of feature importance is useful in facilitating the understanding of a ML model and discovering key features from the dataset. Permutation feature importance[80] is one of the commonly used global interpretability methods which quantifies the importance of a feature as the decrease in terms of model performance after randomly shuffling this feature's value while keeping other features unchanged. Global-level feature importance can also be calculated as the average of all instance-level features' importance derived from methods like SHAP and LIME. For example, the average of the absolute value of SHAP is used as an indicator for global feature importance in the SHAP library.

3.4 Saliency Approach

In the previous sections, I have introduced several interpretable ML methods such as LIME and SHAP that can be used to generate instance-level interpretation of model predictions. These approaches use a surrogate linear model to estimate the causal relationship between input features and model predictions and are also known as feature attribution methods with which an instance-specific prediction can be divided into the contribution of every individual features. In the context of computer vision, the linkage between the features of an image, i.e., a pixel or a group of neighboring pixels, to model predictions can still be established using the methods above.

One way of interpreting a model's prediction back to the image space is to highlight the regions relevant to the model's prediction. In computer vision, the map that highlights different relevant areas is often known as saliency map. If a convolutional neural network predicts "dog" as the classification outcome, a good saliency map, which can be directly overlaid on the original image, should be able to identify the area where there is a dog. In this case, the saliency map is used to assure the prediction of the positive label. However, saliency map can also be used to identify the regions that can help rule out the chance of other categories not being predicted. In this section, I will introduce several broadly used saliency methods and discuss their potential of being used in the medical domain.

One of the most intuitive ways of identifying regions of interest from an image is to observe how the absence of different areas influence model predictions. A model's confidence level of prediction of a certain class might drop significantly if a critical piece of information relevant to that class is masked out. Following this assumption, the strategy of producing a saliency map is by systematically masking out all regions, the drop of predicted probability at distinct locations can thus be stored at the corresponding location. This type of method is known as occlusion- or perturbation-based method[81]. The downside of this approach is the expensive computation cost involved in feed-forwarding thousands of masked images though the model.

Class activation mapping (CAM)[35] is another saliency method that was designed to interpret a convolutional neural network for classification tasks. Because of the ease to use, CAM has been broadly used in many medical image classification problems[82–84] and researchers have developed different variants of CAM[34,85–89] to address the limitations of the original CAM work. As is reported in the original CAM paper, the saliency map can be generated by conducting a weighted summation over all feature maps from the last convolutional layer, where the weights came from the fully connected layer. To apply the CAM method, the authors restricted the CNN to must have a global average pooling layer after the last convolutional layer so that the feature maps can be converted to scaler numbers, which were then passed to the fully connected layer to generate the final prediction. It turns out that the weighted summation of the convolutional feature maps is indeed able to retain the location information of salient features.

As mentioned above, the original CAM methods forced extra restrictions on the model architecture. The Grad-CAM method[34], a variant of the original CAM method, removed the necessity of including global average pooling, and thus can be applied on a wide variety of convolutional neural networks. Like the original version, Grad-CAM also conducts a weighted summation of the features from the last convolutional layer, where the weights were derived from the gradient value through back-propagation. To be more specifically, the gradients on the feature maps from the last convolution block were averaged into a scaler number per channel which is used as the weights for the weight summation over features from all channels.

There are two major limitations of the CAM-based approaches. The first limitation is that the resolution of the CAM-based saliency map is low. In a typical CNN, operations like convolution or pooling which has a stride equals two reduce the spatial dimension of feature maps by a factor of two. Thus, as the feature maps go through the convolutional blocks, the spatial size of the feature maps coming out of the last convolutional layer has shrunk dramatically compared with the original input image size. Because of that, the CAM-based saliency map, as is derived from low resolution feature maps, only shows big blobs on the image which makes it impossible to unveil granular details of the class-specific information. The second limitation of CAM-based saliency approach is rooted from the inexplicit nature of the feature value. Because there is no definite way of interpreting the feature value, there remains a big interpretation gap for the CAM-based saliency map, which is built upon those feature maps, i.e., the interpretation is subject to who is reading the saliency map. Additionally, since there is no fixed scale or range for the saliency value, the interpretation remains at the level of some regions being brighter than other regions. Thus, CAM is not ideal to be applied in the medical domain especially when the disease signature is subtle.

To address the low-resolution issue of the saliency map, a walk-around solution called guided backpropagation[90] was developed to produce the saliency visualization for every pixel of the original image. This method, inspired by the “deconvnet” method[55], altered the rule of gradient backpropagation by only allowing the backflow of positive gradients. Though this approach can generate high resolution saliency maps, the produced saliency value is not class-specific, i.e., highlighted regions could be relevant to any categories of the classification task. Later, Selvaraju et al. combined the class-specific Grad-CAM saliency map with that from the guided backpropagation by simply multiplying these two types of saliency maps[34].

Though we have witnessed rapid progress toward the development of saliency methods on natural images, there are still urgent needs for medicine-specific saliency methods. In this dissertation, I present a novel saliency method called disease probability map in chapter 4 and a SHAP-based saliency analysis is reported in chapter 5. Both methods can produce high resolution saliency map to reveal subtle disease signatures of AD and dementia, but the underlying mechanism of these two methods are dramatically different. Each of these approaches satisfied some properties suitable for the applications in medical domain. The quality of the saliency map can normally be inspected by observing the correspondence between the highlighted areas and true object locations for natural images. However, the same strategy may not be feasible for the medical problems, e.g., predicting pneumonia from a chest CT scan and diagnosing Alzheimer's disease from a brain MRI scan, due to the lack of domain knowledge. In this dissertation, I will also present the systematic way we used to validate these saliency methods.

Chapter 4

Interpretable Deep Learning for Alzheimer's Disease

4.1 Introduction

Millions worldwide continue to suffer from Alzheimer's disease (AD), while attempts to effectively treat the disease remain stalled. Though tremendous progress has been made towards detecting AD pathology using CSF biomarkers[91], as well as PET amyloid[92], and tau imaging[93], these modalities often remain limited to research contexts. Instead, current standards of diagnosis depend on highly skilled neurologists to conduct an examination that includes inquiry of patient history, an objective cognitive assessment such as Mini-Mental State Examination (MMSE) or neuropsychological testing[94], and a structural MRI to rule in findings suggestive of AD[95]. Clinicopathological studies suggest the diagnostic sensitivity of clinicians ranges between 70.9% and 87.3% and specificity between 44.3% and 70.8%[96]. Given this relatively imprecise diagnostic landscape, as well as the invasive nature of CSF and PET diagnostics and a paucity of clinicians with sufficient Alzheimer's disease diagnostic expertise, advanced machine learning paradigms such as deep learning offer ways to derive high accuracy predictions from MRI data and other measures collected within the bounds of neurology practice.

Part of this chapter is reprinted from "Development and validation of an interpretable deep learning framework for Alzheimer's disease classification." By Qiu, Shangran, Prajakta S. Joshi, Matthew I. Miller, Chonghua Xue, Xiao Zhou, Cody Karjadi, Gary H. Chang et al. *Brain* 143, no. 6 (2020): 1920-1933. Copyright © 2022 Oxford University Press.

Recent studies have demonstrated the application of deep learning models such as CNN for MRI and multimodal data-based classification of cognitive status[97]. These models have yet to achieve full integration into clinical practice for several reasons. First, there is a lack of external validation of deep learning algorithms since most models are trained and tested on a single cohort. Second, there is a growing notion in the biomedical community that deep learning models are 'black-box' algorithms[98]. In other words, although deep learning models demonstrate high accuracy classification across a broad spectrum of disease, they neither elucidate the underlying diagnostic decisions nor indicate the input features associated with the output predictions. Thus, overcoming these challenges is not only crucial to harness the potential of deep learning algorithms to improve patient care, but to also pave the way for explainable evidence-based machine learning in the medical imaging community. In this chapter, a novel deep learning framework that we developed will be presented which links a fully convolutional network (FCN) to a traditional MLP to generate high-resolution visualizations of Alzheimer's disease risk that can then be used for accurate predictions of Alzheimer's disease status. This framework was validated using four distinct cohorts along with neuropathological findings and a head-to-head comparison with a team of neurologists.

4.2 Dataset

4.2.1 Dataset Collection

Data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL), the Framingham Heart Study (FHS), and the National Alzheimer's Coordinating Center (NACC) cohorts were used in the study. ADNI is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of AD[99]. AIBL, launched in 2006, is the largest study of its kind in Australia and aims to discover biomarkers, cognitive characteristics, and lifestyle factors that influence the development of symptomatic AD[100]. The FHS is a longitudinal community cohort study and has collected a broad spectrum of clinical data from three generations[101]. Since 1976, the FHS expanded to evaluate factors contributing to cognitive decline, dementia, and AD. Finally, the NACC, established in 1999, maintains a large relational database of standardized clinical and neuropathological research data collected from AD centers across the USA[102].

Model training, internal validation and testing were performed on the ADNI dataset. Following training and internal testing on the ADNI data, we validated the predictions on AIBL, FHS, and NACC. The criterion for selection included individuals aged ≥ 55 years, with 1.5 Tesla, T1-weighted MRI scans taken within ±6 months from the date of clinically confirmed diagnosis of Alzheimer's disease or normal cognition. We excluded cases including Alzheimer's disease with mixed dementia, non-Alzheimer's disease dementias, history of severe traumatic brain injury, severe depression, stroke, and brain tumors, as

Dataset	ADNI			AIBL			FHS			NACC		
Characteristic	**NC (n=229)**	**AD (n=188)**	**p-value**	**NC (n=320)**	**AD (n=62)**	**p-value**	**NC (n=73)**	**AD (n=29)**	**p-value**	**NC (n=356)**	**AD (n=209)**	**p-value**
Age, y, median [range]	76 [60, 90]	76 [55, 91]	0.4185	72 [60, 92]	73 [55, 93]	0.5395	73 [57, 100]	81 [67, 94]	<0.0001	74 [56, 94]	77 [55, 95]	0.0332
Education, y, median [range]	16 [6, 20]	16 [4, 20]	<0.0001	N.A.	N.A.	N.A.	14 [8, 25]	13 [5, 25]	0.3835	16 *a [0, 22]	14.5 *b [2, 24]	0.8363
Gender, male (%)	119 (51.96)	101 (53.72)	0.7677	144 (45.00)	24 (38.71)	0.4031	37 (50.68)	12 (41.38)	0.5105	126 (35.39)	95 (45.45)	0.0203
MMSE, median [range]	29 [25, 30]	23.5 [18, 28]	<0.0001	29 [25, 30]	21 [6, 28]	<0.0001	29 *e [22, 30]	25 [10, 29]	<0.0001	29 *c [20, 30]	22 *d [0:30]	<0.0001
APOE4, positive (%)	61 (27)	124 (66)	<0.0001	11 (3.4)	12 (19.4)	<0.0001	13 (17.81)	11 *f (40.74)	0.0355	102 (28.65)	112 (53.59)	<0.0001

Table 4.1. Study population and characteristics. Four independent datasets were used for this study including: the ADNI dataset, the AIBL, the FHS, and the NACC. The ADNI dataset was randomly split in the ratio of 3:1:1, where 60% of it was used for model training, 20% of the data was used for internal validation and the rest was used for internal testing. The best performing model on the validation dataset was selected for making predictions on the ADNI test data as well as on the AIBL, FHS and NACC datasets, which served as external test datasets for model validation. All the MRI scans considered for this study were performed on individuals within ±6 months from the date of clinical diagnosis. AD = Alzheimer's disease; NA = not available; NC = normal cognition.

*a -- Education information not available for two subjects

*b -- Education information not available for one subject

*c -- MMSE score was not available for one subject

*d -- MMSE score was not available for one subject

*e -- Six subjects do not have available MMSE scores

*f -- APOE4 information not available for one subject well as incident major systemic illnesses. Note that this inclusion and exclusion criterion was adapted from the baseline recruitment protocol developed by the ADNI study99, and to maintain consistency, the same criterion was applied to other cohorts as applicable. This led to the selection of 417 individuals from the ADNI cohort, 382 individuals from AIBL, 102 FHS participants, and 565 individuals from the NACC cohort (Table 4.1). If an individual had multiple MRI scans taken within the time window, then we selected the scan closest to the date of clinical diagnosis. For most of these selected cases, age, gender and MMSE score were available.

4.2.2 Data Harmonization

The MRI scans from all the datasets were obtained in NIFTI format. We used the MNI152 template (ICBM 2009c Nonlinear Symmetric template, McGill University, Canada) to register all the scans. We used the FLIRT tool available within the FSL package (Wellcome Center, University of Oxford, UK), to align the scans with respect to the MNI152 template. A careful manual review of the registered images revealed that the automatic registration was done reasonably well on a large majority of the ADNI, AIBL and NACC cases. For cases that were not registered well (mainly within FHS), we performed affine transformations to perform manual registration using known regions as landmarks. Given that there may not be a registration method that would work for all MRI scans, our two-step process resulted in a reasonable set of registered images. After image registration, we normalized intensities of all the voxels [mean = 0 and standard deviation (SD) = 1]. We then adjusted the intensity of these voxels and other outliers by clipping them to the range: [-1, 2.5], where any voxel with intensity lower than -1 was assigned a value of −1, and a voxel with intensity higher than 2.5 was assigned a value of 2.5. We then performed background removal where all the voxels from background regions outside of the skull were set to -1 to ensure uniform background intensity.

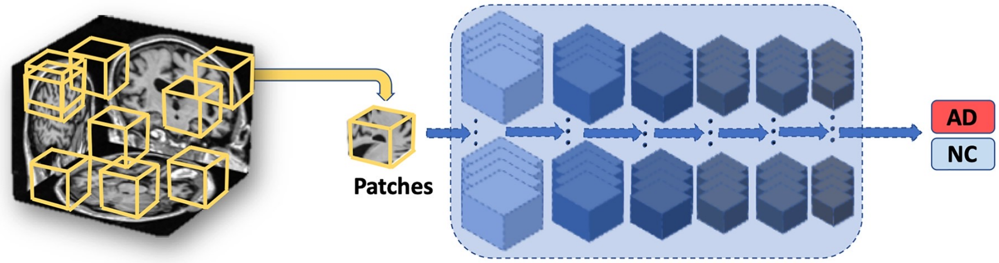
4.3 Model Development

4.3.1 Saliency Method

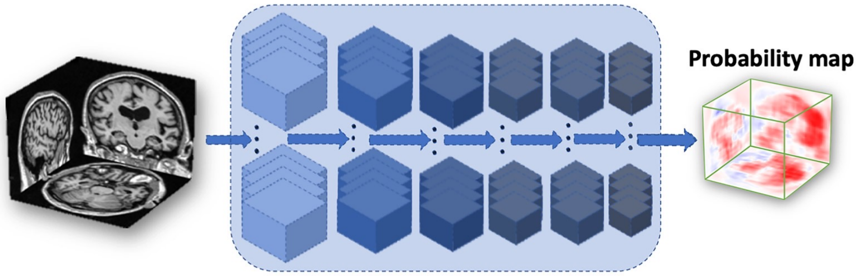
An FCN was designed to input a registered volumetric MRI scan of size $181 \times 217 \times 181$ voxels and output the Alzheimer's disease class probability at every location. We used a novel, computationally efficient patch-wise training strategy to train the FCN model (Figure 4.1). This process involved random sampling of 3000 volumetric patches of size $47 \times 47 \times 47$ voxels from each training subject's MRI scan and used this information to predict the output of interest. The size of the patches was the same as the receptive field of the FCN.

The FCN consists of six convolutional blocks. The first four convolutional blocks consist of a 3D convolutional layer followed by the following operations: 3D max pooling, 3D batch-normalization, Leaky ReLu and Dropout. The last two convolutional layers function as dense layers in terms of the classification task and these two layers play a key role in boosting model efficiency[57]. The network was trained *de novo* with random initialization of the weights. We used the Adam optimizer with a 0.0001 learning rate and a mini-batch size of 10. During the training process, the model was saved when it achieved the lowest error on the ADNI validation dataset.

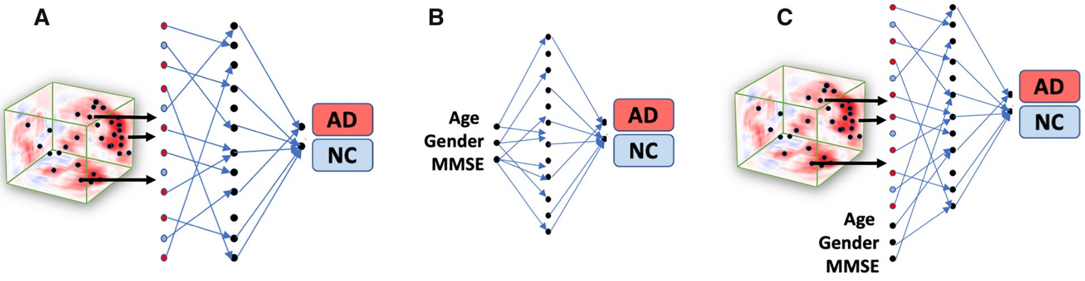
The FCN was trained by repeated application to cuboidal patches of voxels randomly sampled from a full volume of sequential MRI slices. Because the convolutions decrease the size of the input across successive layers of the network, the size of each patch was selected such that the shape of the final output from each patch was equal to $2 \times 1 \times 1 \times 1$; i.e., the application of the FCN to each patch during training produced a list of two scalar values. These values can be converted to respective Alzheimer's disease and normal cognition probabilities by application of a SoftMax function. In this way, the model was trained to infer local patterns of cerebral structure that suggested an overall disease state.



Step 1: Random sampling of patches for fully convolutional network training



Step 2: Generate probability maps after fully convolutional network training



Step 3: Multilayer perceptrons generate overall prediction.

Figure 4.1. Schematic of the deep learning framework. The FCN model was developed using a patch-based strategy in which randomly selected samples of T1-weighted full MRI volumes were passed to the model for training (Step 1). The corresponding Alzheimer's disease status of the individual served as the output for the classification model. Given that the operation of FCNs is independent of input data size, the model led to the generation of participant-specific disease probability maps of the brain (Step 2). Selected voxels of high-risk from the disease probability maps were then passed to the MLP for binary classification of disease status (Model A in Step 3; MRI model). As a further control, we used only the non-imaging features including age, gender and MMSE and developed an MLP model (Model B in Step 3; non-imaging model). We also developed another model that integrated multimodal input data including the selected voxels of high-risk disease probability maps alongside age, gender and MMSE score to perform binary classification of Alzheimer's disease status (Model C in Step 3; Fusion model).

4.3.2 Disease Probability Map

After the training process of the FCN model, registered whole volumetric MRI scans of size $181 \times 217 \times 181$ voxels were sent into the FCN model to output the Alzheimer's disease class probability at every location. We refer the output of the FCN model as disease probability map (DPM). For illustration, the DPMS from 4 randomly selected individuals were presented in Figure 4.2 as colored heatmaps where red regions indicate high risk of AD and blue spots indicates low risk of AD. Because the voxel value represents the probability that the subject has AD, all voxel values range from zero to one. The process of obtaining disease probability maps from test cases took ∼1 second on an NVIDIA GTX Titan GPU.

4.3.3 Multimodal Data Integration

After generating disease probability maps over all subjects, an MLP model framework was developed to perform binary classification to predict Alzheimer's disease status by selecting Alzheimer's disease probability values from the DPMs. This selection was based on observation of the overall performance of the FCN classifier as estimated using the Matthew's correlation coefficient values on the ADNI training data. Specifically, we selected DPM voxels from 200 fixed locations that were indicated to have high Matthew's correlation coefficient (MCC) values.

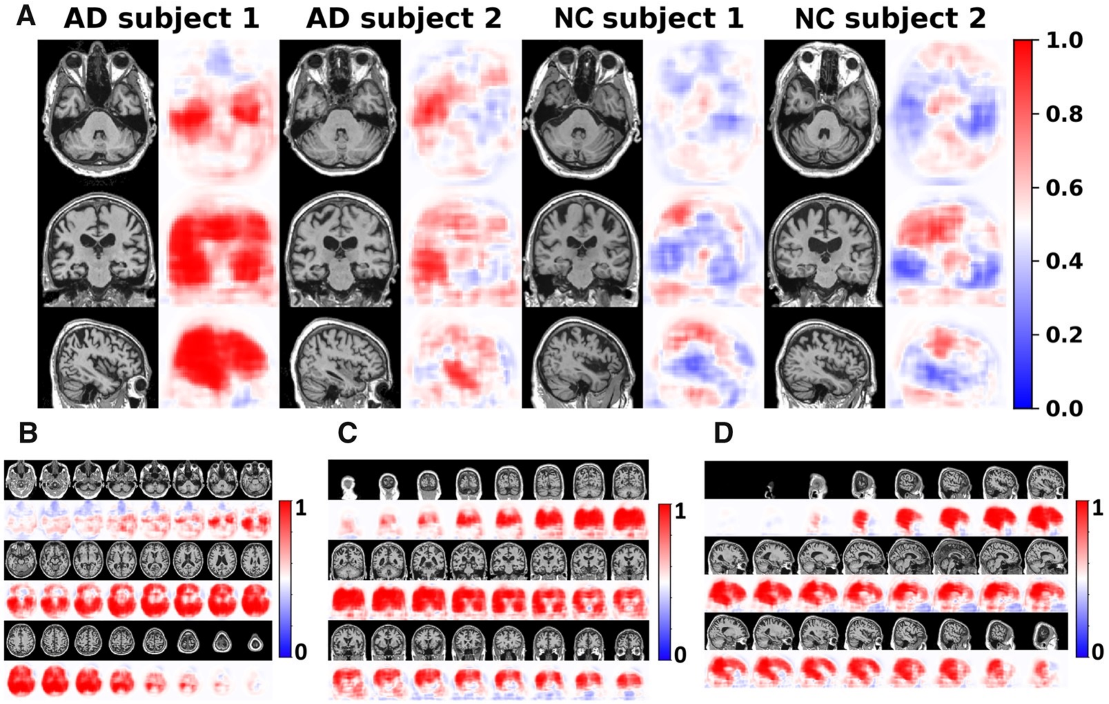


Figure 4.2. Subject-specific disease probability maps. (A) Disease probability maps generated by the FCN model highlight high-risk brain regions that are associated with AD pathology. Individual cases are shown where the blue color indicates low-risk and red indicates high-risk of AD. The first two individuals had a clinical diagnosis of AD whereas the other two were clinically confirmed to have normal cognition. (B–D) Axial, coronal, and sagittal stacks of disease probability maps from a single subject with clinically confirmed Alzheimer's disease are shown. All imaging planes were used to construct 3D disease probability maps. Red color indicates locally inferred probability of Alzheimer's disease > 0.5, whereas blue indicates < 0.5.

The features extracted from these locations served as input to the MLP model that performed binary classification of Alzheimer's disease status (MRI model in Figure 4.1, Step 3). Two additional MLP models were developed where one model used age, gender, and MMSE score values as input to predict Alzheimer's disease status (non-imaging model in Figure 4.1, Step 3), and the other MLP took the 200 features along with age, gender, and MMSE score as input to predict Alzheimer's disease status (Fusion model in Figure 4.1, Step 3). All the MLP models comprised a single hidden layer and an output layer. The MLP models also included non-linear operators such as ReLu and Dropout.

4.4 Validation

4.4.1 Local Accuracy

To assess the agreement between the individuals' diagnostic labels and the voxel values from the DPMs, we derived the population-wide maps of MCC (Figure 4.3) which is a commonly used performance metric for classification task and its value ranges from -1 to 1. MCC is generally considered as a balanced metric which captures the quality of the classification even if data is imbalance across classes. By treating each location independently, MCC value was derived based on a list of diagnostic labels and corresponding probabilities from each spatial location. Thus, the maps of MCC characterize the distribution of local "accuracy" of the DPMs. This mapping enabled identification of areas from which correct predictions of disease status were most frequently derived, thus acting to demonstrate structures most affected by neuropathological changes in Alzheimer's disease.

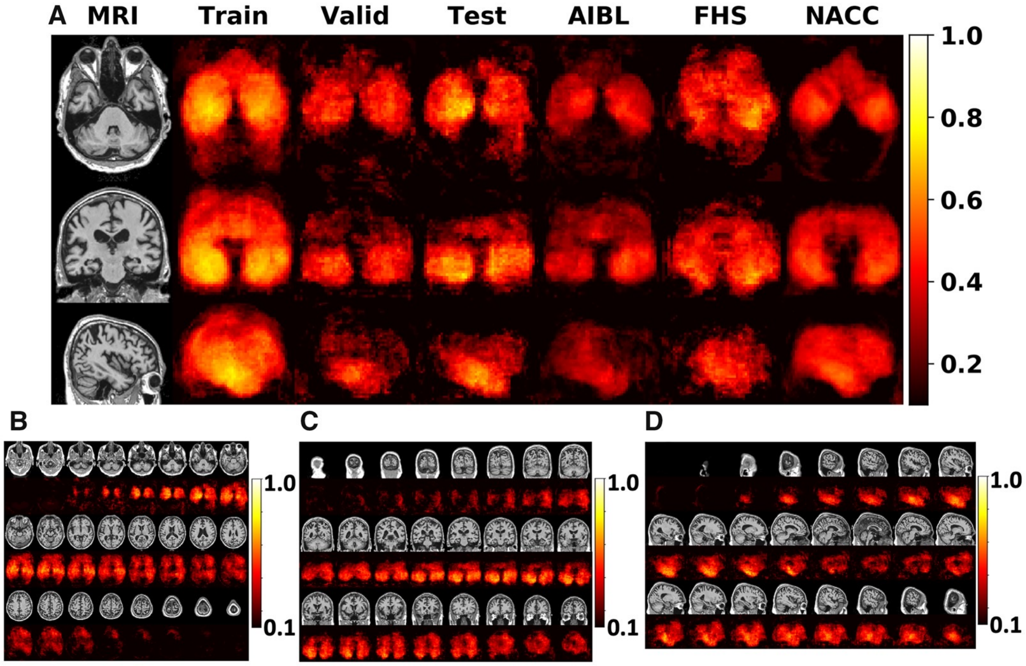


Figure 4.3. Local accuracy of the FCN model performance. (A) Voxel-wise maps of MCC were computed independently across all the datasets to demonstrate predictive performance derived from all regions within the brain. (B–D) Axial, coronal, and sagittal stacks of the MCC maps at each cross-section from a single subject, are shown. These maps were generated by averaging the MCC values on the ADNI test data.

4.4.2 Neuropathological Validation

As confirmation, average regional probabilities extracted from selected segmented brain regions were highly associated with Alzheimer's disease positive findings reported in post-mortem neuropathology examinations (Figure 4.4). Specifically, these regions correlated with the locations and numerical frequency of amyloid-β, and tau pathologies reported in available autopsy reports from the FHS dataset (n = 11). Post-mortem data indicated that, in addition to predicting higher region-specific Alzheimer's disease probabilities in individuals with disease compared to those without, proteinopathies were more frequent in cerebral regions implicated by the model in Alzheimer's disease (Figure 4.4).

Model-predicted regions of high Alzheimer's disease risk overlapped with the segmented regions that were indicated to have high localized deposition of amyloid-β and tau. Additionally, predicted Alzheimer's disease risk within these zones increased with pathology scores. Given that these post-mortem findings are definitive in terms of confirming Alzheimer's disease, these physical findings grounded our computational predictions in biological evidence.

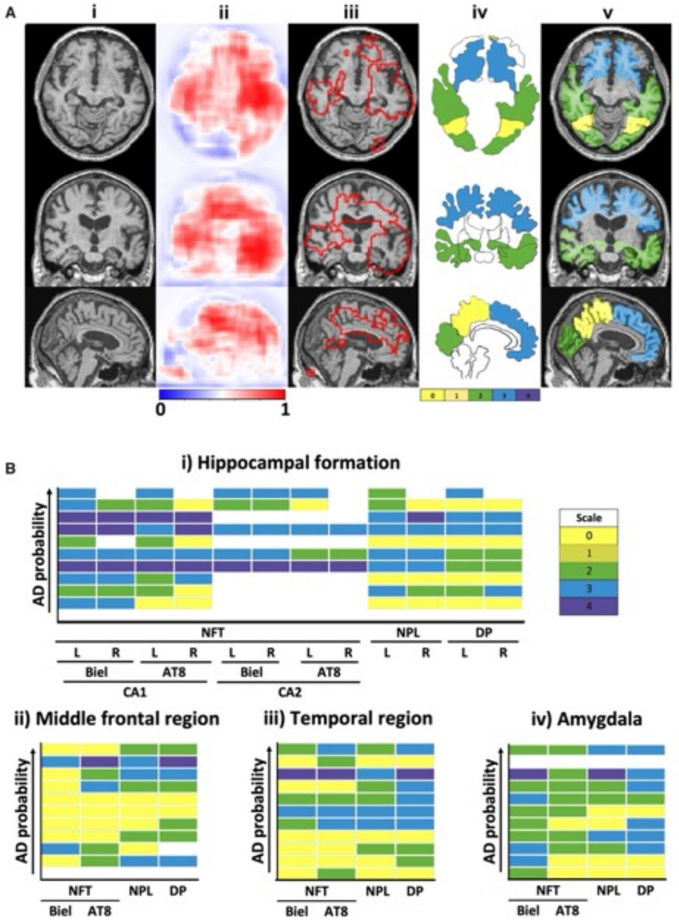


Figure 4.4. Correlation of model findings with neuropathology. (A) Overlap of model predicted regions of high Alzheimer's disease risk with post-mortem findings of Alzheimer's disease pathology in a single subject. This subject had clinically confirmed Alzheimer's disease with affected regions including the bilateral asymmetrical temporal lobes and the right-side hippocampus, the cingulate cortex, the corpus callosum, part of the parietal lobe and the frontal lobe. The first column (i) shows MRI slices in three different planes followed by a column (ii), which shows corresponding model predicted disease probability maps. A cut-off value of 0.7 was chosen to delineate the regions of high Alzheimer's disease risk and overlapped with the MRI scan in the next column (iii). The next column (iv) depicts a segmented mask of cortical and subcortical structures of the brain obtained from FreeSurfer. A sequential color-coding scheme denotes different levels of pathology ranging from green (0, low) to pale red (4, high). The final column (v) shows the overlay of the magnetic resonance scan, disease probability maps of high Alzheimer's disease risk and the color-coded regions based on pathology grade. (B) We then qualitatively assessed trends of neuropathological findings from the FHS dataset (n = 11). The same color-coding scheme as described above was used to represent the pathology grade (0–4) in the heat maps. The boxes colored in 'white' in the heat maps indicate missing data. Using the Spearman's Rank correlation coefficient test, an increasing Alzheimer's disease probability risk was associated with a higher grade of amyloid-β and tau accumulation, in the hippocampal formation, the middle frontal region, the amygdala and the temporal region, respectively.

4.4.3 Global Accuracy Against Neurologists

As confirmation, average regional probabilities extracted from selected segmented brain regions disease probability maps provided an information-dense feature that yielded sensitive and specific binary predictions of Alzheimer's disease status when passed independently to the MLP portion of the framework (MRI model in Figure 4.5. a & b). An MLP trained using just the non-imaging features such as age, gender, and MMSE score also was predictive of Alzheimer's disease status (non-imaging model in Figure 4.5. a & b). Model performance was further improved by expanding the MLP input to include DPMs, gender, age, and MMSE score (fusion model in Figure 4.5. a & b). When other non-imaging features such as APOE status were included, model performance slightly improved. Given the proportionality between age and global cerebral atrophy[103,104], addition of non-imaging variables at the MLP stage also allowed us to control for the natural progression of cerebral morphological changes over the lifespan.

We also compared performance of the deep learning models against an international group of clinical neurologists recruited to provide impressions of disease status from a randomly sampled cohort of ADNI participants whose MRI, MMSE score, age, and gender were provided. The performance of the neurologists (Figure 4.5. a) indicated variability across different clinical practices, with a moderate inter-rater agreement as assessed by pairwise kappa (κ) scoring (Figure 4.5. a; average $\kappa = 0.493 \pm 0.16$). Interestingly, we noted that the deep learning model that was based on MRI data alone (MRI model; accuracy: 0.834 ± 0.020), outperformed the average neurologist (accuracy: 0.823 ± 0.094).

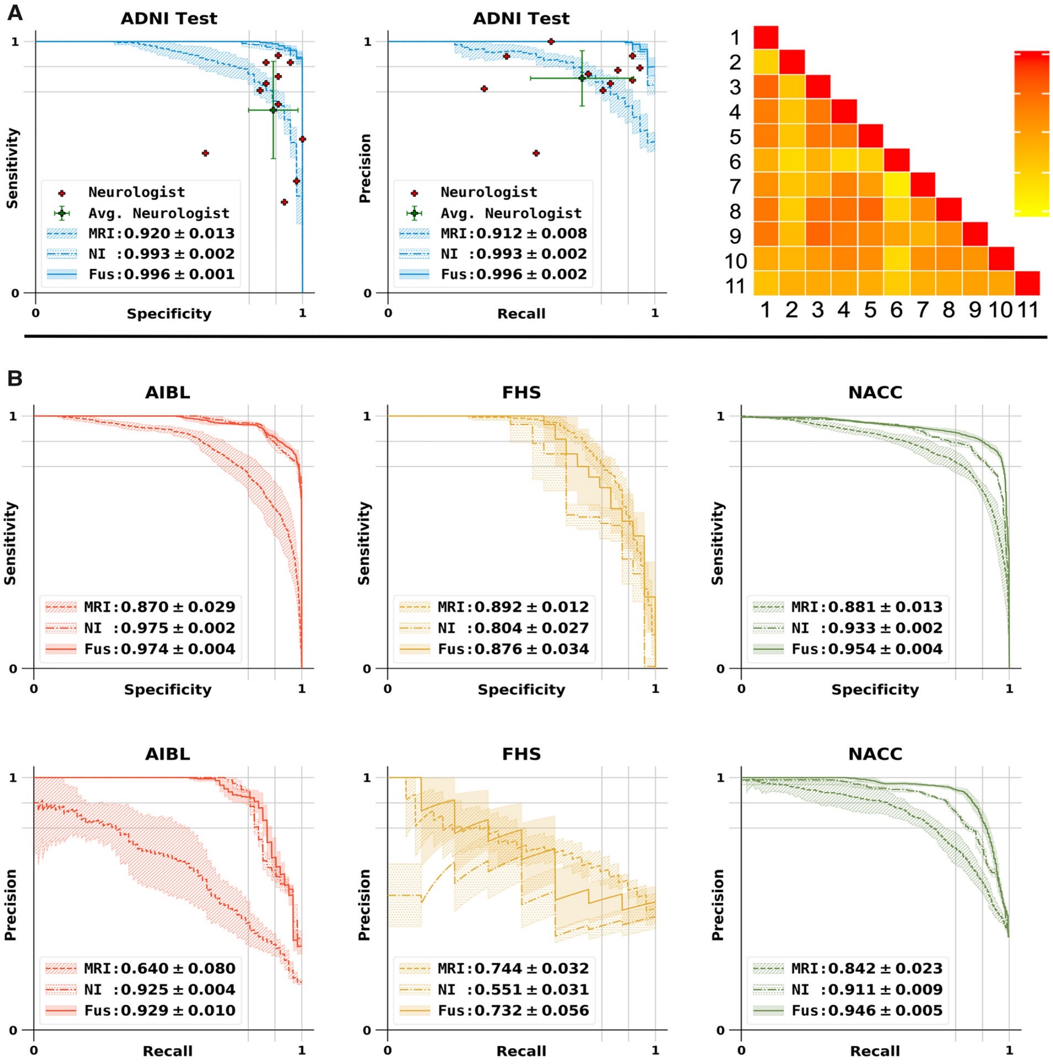


Figure 4.5. Performance of the MLP model for Alzheimer's disease classification and model comparison with neurologists. (A) Sensitivity-specificity and precision-recall curves showing the sensitivity, the true positive rate, versus specificity, the true negative rate, calculated on the ADNI test set. Individual neurologist performance is indicated by the red plus symbol and averaged neurologist performance along with the error bars is indicated by the green plus symbol on both the sensitivity-specificity and precision-recall curves on the ADNI test data. Visual description of pairwise Cohen's kappa (κ), which denotes the inter-operator agreement between all the 11 neurologists is also shown. (B) Sensitivity-specificity and PR curves calculated on the AIBL, FHS and NACC datasets, respectively.

When age, gender and MMSE information were added to the model, then the performance increased significantly (fusion model; accuracy: 0.968 ± 0.014).

Consistent, high classification performance of the deep learning model across the external datasets was confirmed using other metrics. We performed t-distributed stochastic neighbor embedding (t-SNE)[105], on the volumetric MRI scans using the intensity values as inputs from all the four datasets. The t-SNE method takes high-dimensional data and creates a low-dimensional representation of those data, so that it can be easily visualized. While the t-SNE plot resulted in site-specific clustering of the scans (Figure 4.6. a), intra-site distribution of cases revealed no clear differentiation between Alzheimer's disease and normal cognition cases.

This observation underscores a rationale for utilizing a supervised learning strategy to predict Alzheimer's disease status using MRI scan data alone. We believe this is a strength of our study because despite site-specific differences, the FCN model was able to generalize well on the external datasets. We then used scanner-specific info from the ADNI cohort and generated another t-SNE visualization, which also revealed no discernible clustering of Alzheimer's disease or normal cognition cases (Figure 4.6. b). This implies that any potential scanner-specific differences may not have influenced the model training process. Further, we examined the model performance visually by respective clustering of Alzheimer's disease and normal cognition cases in a t-SNE, which used features before the final hidden layer of the MLP (Figure 4.6. c).

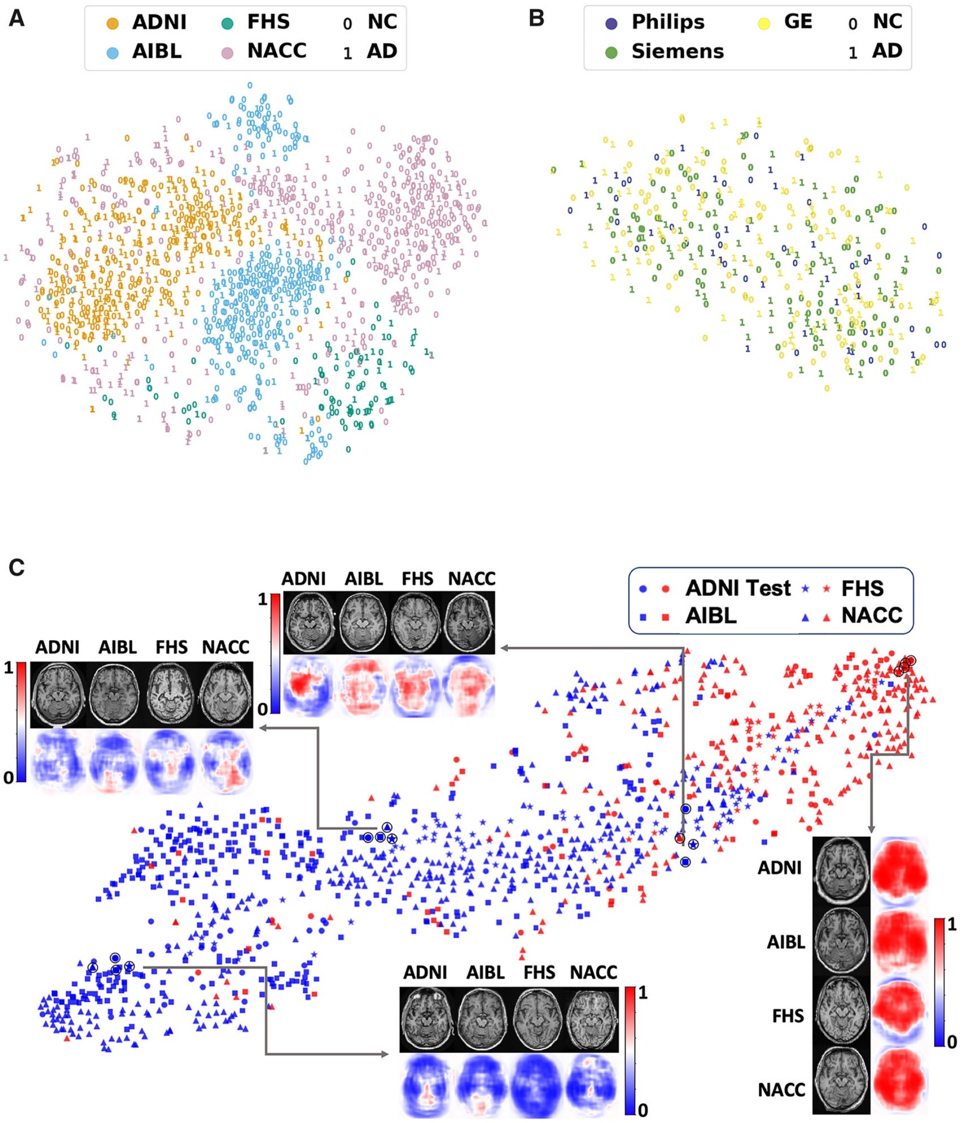


Figure 4.6. Visualization of data. (A) Voxel-level MRI intensity values from all four datasets (ADNI, AIBL, FHS and NACC) were used as inputs for a t-SNE plot. The color in the plot represents the site and the marker represents the label. (B) This t-SNE plot was generated only on using the ADNI dataset, where the color was used to represent the scanner. (C) FCN-based outputs that served as input features to the MLP model were embedded in a two-dimensional plot generated using t-SNE for the two classes (AD and NC). The color (blue versus red) was used to distinguish normal cognition from AD cases, whereas a unique symbol shape was used to represent individuals derived from the same cohort.

4.5 Conclusion

Our deep learning framework links an FCN to a MLP and generates high resolution DPMs for neurologist-level diagnostic accuracy of AD status. The intuitive local probabilities outputted by our model are readily interpretable, thus contributing to the growing movement towards explainable artificial intelligence in medicine and deriving an individualized phenotype of insidious disease from conventional diagnostic data. Indeed, the DPMs provide a means for tracking conspicuous brain regions implicated in AD. We then aggregated DPMs across the entire cohort to demonstrate population-level differences in neuroanatomical risk mapping of AD and normal cognition cases. Critically, by the standards of several different metrics, our model displayed good predictive performance, yielding high and consistent values on all the test datasets. Such consistency between cohorts featuring broad variance in MRI protocol, geographic location, and recruitment criteria, suggests a strong degree of generalizability. Thus, these findings demonstrate innovation at the nexus of medicine and computing, simultaneously contributing new insights to the field of computer vision while also expanding the scope of biomedical applications of neural networks.

Certainly, limitations to this study must be acknowledged. We considered a case-control population in which two subpopulations were chosen in advance that were either cognitively normal or have the diagnosis of AD. This scenario is not exactly representative of the standard clinical decision-making process faced by the neurologist. Patients often present with a set of symptoms and results from standard neurological testing that are indicative of a spectrum of neurodegenerative disease as opposed to a binary scenario.

Therefore, our method is not directly applicable in its current state but serves as a first step towards building a more comprehensive framework to characterize multiple etiologies of neurodegeneration.

Our approach has significant translational potential beyond AD diagnosis. Indeed, the tissue-level changes predicted by our model suggest the prospect of directly highlighting areas of pathophysiology across a spectrum of disease. It may be of interest in future studies to determine whether the well-defined pattern of high-risk findings from currently presented framework may follow regions of interest from PET scans. In such cases, our model may aid in non-invasive monitoring of AD development.

In conclusion, our deep learning framework was able to obtain high accuracy AD classification signatures from MRI data, and our model was validated against data from independent cohorts, neuropathological findings, and expert-driven assessment. If confirmed in clinical settings, this approach has the potential to expand the scope of neuroimaging techniques for disease detection and management. Further validation could lead to improved care and outcomes compared with current neurological assessment, as the search for disease-modifying therapies continues.

Chapter 5

Expert-level Deep Learning for Dementia Assessment

5.1 Introduction

Despite the progress towards advances of AD biomarkers and novel imaging scans like tau positron emission tomography (PET), these modalities remain limited to research contexts, and the backbone of diagnosis still rely on traditional clinical assessment. Mild cognitive impairment (MCI), a prodromal stage of dementia, may also be a subtle early presentation of AD whose diagnosis similarly requires significant clinical acumen from qualified specialists. Complicating matters is the presence of multiple dementia etiologies in a subject, such as AD, vascular dementia (VD), Lewy body dementia (LBD), and frontotemporal dementia (FTD), metabolic disorder, traumatic injury, infectious disease etc., which widen the differential diagnosis of neurodegenerative conditions and contribute to variability in diagnostic sensitivity and specificity[96]. As the age of the subject increases, the likelihood of observing multiple dementia etiologies within the subject also increases. Thus, many of the individuals who were diagnosed with AD also have other dementia commodities.

Reliably differentiating between normal cognitive aging, MCI, AD, and other dementia etiologies requires significant clinical acumen from qualified specialists treating

---

Part of this chapter is reprinted from "Multimodal deep learning for Alzheimer's disease dementia assessment." By Qiu, Shangran, Matthew I. Miller, Prajakta S. Joshi, Joyce C. Lee, Chonghua Xue, Yunruo Ni, Yuwei Wang, et al. *Nature Communications* 13, 3404 (2022) Copyright © 2022 Springer Nature.

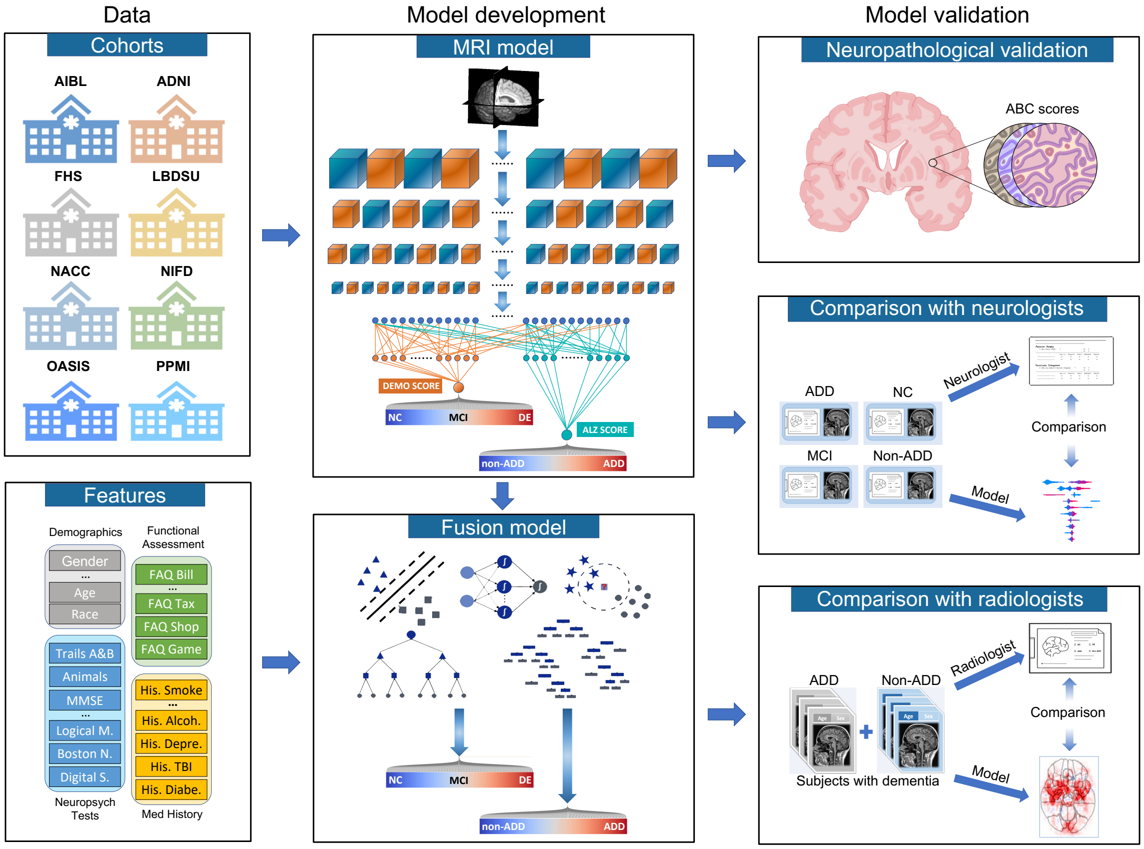


Figure 5.1. Study design of a multi-task deep learning framework. Multimodal data including MRI scans, demographics, medical history, functional assessments, and neuropsychological test results were used to develop deep learning models on various classification tasks. Eight independent datasets were used for this study, including NACC, ADNI, AIBL, FHS, LBDSU, NIFD, OASIS, and PPMI. We selected the NACC dataset to develop three separate models: (i) an MRI-only CNN model (ii) non-imaging models in the form of traditional machine learning classifiers, which did not use any MRI data (iii) a fusion model combining imaging and non-imaging data within a hybrid architecture joining a CNN to a CatBoost model. First, T1-weighted MRI scans were input to a CNN to calculate a continuous DEmentia MOdel (DEMO) score to assess cognitive status on a 0 to 2 scale. For individuals with DE diagnosis, the multi-task CNN model simultaneously discriminated their risk of having AD versus nADD, a classification that we refer to as the ADD task. We denoted the probability of AD diagnosis as the ALZheimer (ALZ) score.

memory disorders, yet timely access to memory clinics is often limited for patients and families. There is a dearth of specialized practitioners globally. Furthermore, the need for skilled clinicians is rising, yet the United States is facing a projected shortage of qualified clinicians, such as neurologists, in coming decades[106,107]. As increasing clinical demand intersects with a diminishing supply of medical expertise, machine learning methods for aiding neurologic diagnoses have begun to attract interest. Complementing the high diagnostic accuracy reported by other groups[108], we reported interpretable deep learning approaches[37] capable of distinguishing participants with age-appropriate normal cognition (NC) from those with AD using magnetic resonance imaging (MRI) scans, age, sex, and mini-mental state examination (MMSE) in chapter 4. Others have also demonstrated the efficacy of deep learning in discriminating AD from specific types of nADD[109,110]. However, clinical evaluation of persons presenting in a memory clinic involves consideration of multiple etiologies of cognitive impairment. Therefore, the ability to successfully differentiate between NC, MCI, AD, and nADD across diverse study cohorts in a unified framework remains to be developed.

Here we report a deep learning framework that accomplishes multiple diagnostic steps in successive fashion to identify persons with normal cognition (NC), mild cognitive impairment (MCI), AD, and non-AD dementias (nADD). We demonstrate a range of models capable of accepting flexible combinations of routinely collected clinical information, including demographics, medical history, neuropsychological testing, neuroimaging, and functional assessments. We then show that these frameworks compare favorably with the diagnostic accuracy of practicing neurologists and neuroradiologists.

Lastly, we apply interpretability methods in computer vision to show that disease-specific patterns detected by our models track distinct patterns of degenerative changes throughout the brain and correspond closely with the presence of neuropathological lesions on autopsy. Our work demonstrates methodologies for validating computational predictions with established standards of medical diagnosis (Figure 5.1).

5.2 Dataset

5.2.1 Dataset Collection and Feature Selection

We collected demographics, medical history, neuropsychological tests, and functional assessments as well as magnetic resonance imaging (MRI) scans from 8 cohorts (Table 5.1), totaling 8,916 participants after assessing for inclusion criteria. There were 4,550 participants with normal cognition (NC), 2,412 participants with mild cognitive impairment (MCI), 1,606 participants with Alzheimer's disease dementia (AD) and 348 participants with dementia due to other causes. The eight cohorts include the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (n=1,821), the National Alzheimer's Coordinating Center (NACC) dataset (n=4,822), the frontotemporal lobar degeneration neuroimaging initiative (NIFD) dataset (n=253)[111], the Parkinson's Progression Marker Initiative (PPMI) dataset (n=198)[112], the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL) dataset (n=661)[113], the Open Access Series of Imaging Studies-3 (OASIS) dataset (n=666)[114], the Framingham Heart Study (FHS) dataset (n=313)[115], and in-house data maintained by the Lewy Body Dementia Center for Excellence at Stanford University (LBDSU) (n=182)[116].

Dataset (group) [subjects]		Age Mean ±std	Gender Male (percent)	Education In years Mean ±std	Race (White; Black; Asian; Indian; Pacific; Multi-race)	ApoE4 Positive (percent)	MMSE Mean ±std	CDR Mean ±std	MOCA Mean ±std
ADNI	NC [n=481]	74.26 ±6.00	235 (48.86%)	16.34 ±2.67	(436, 36, 8, 1, 0, 3)^	138 (29.61%) ^	29.05 ±1.12	0.00 ±0.00^	25.71 ±2.59^
	MCI [n=971]	72.84 ±7.71	572 (58.91%)	15.94 ±2.81	(903, 34, 15, 2, 2, 12)^	438 (47.20%) ^	27.62 ±1.81	0.50 ±0.04	23.18 ±3.23^
	AD [n=369]	74.91 ±7.84	203 (55.01%)	15.18 ±2.97	(343, 15, 7, 0, 0, 4)	229 (64.33%) ^	23.19 ±2.11	0.77 ±0.26	16.80 ±4.50^
	p-value	2.565e-6	1.364e-3	1.872e-8	1.132e-1	3.117e-22	<1.0e-200	<1.0e-200	1.010e-116
NACC	NC [n=2524]	69.82 ±9.93^	871 (34.51%)	15.92 ±2.95^	(2120, 303, 55, 31, 2, 0)^	599 (29.95%) ^	28.98 ±1.31^	0.06 ±0.16^	26.80 ±2.44^
	MCI [n=1175]	74.01 ±8.74^	555 (47.23%)	15.36 ±3.35^	(965, 160, 25, 17, 1, 0)^	322 (38.66%) ^	26.79 ±2.51^	0.46 ±0.18^	22.68 ±3.41^
	AD [n=948]	74.97 ±9.13^	431 (45.46%)	14.64 ±3.64^	(816, 85, 23, 11, 0, 0)^	346 (52.19%) ^	20.48 ±5.69^	1.02 ±0.60^	15.39 ±5.44^
	Non-AD [n=175]	69.35 ±10.84^	110 (62.86%)	14.86 ±3.60^	(161, 10, 2, 1, 0, 0)^	34 (25.95%) ^	22.23 ±6.14^	1.07 ±0.70^	17.53 ±6.35^
	p-value	1.145e-56	1.130e-22	1.846e-25	5.349e-2	8.026e-49	<1.0e-200	<1.0e-200	<1.0e-200
NIFD	NC [n=124]	63.21 ±7.27	56 (45.16%)	17.48 ±1.87^	(89, 0, 0, 0, 0, 3)^	N.A.	29.35 ±0.76	0.03 ±0.12^	27.58 ±1.53^
	Non-AD [n=129]	63.66 ±7.33	75 (58.14%)	16.18 ±3.29^	(109, 1, 1, 0, 0, 4)^	N.A.	24.75 ±4.54^	0.82 ±0.54^	19.69 ±5.72^
	p-value	6.266e-1	5.246e-2	2.606e-4	6.531e-1	N.A.	1.961e-23	4.333e-28	2.645e-16
PPMI	NC [n=171]	62.74 ±10.12	109 (63.74%)	15.82 ±2.93	(163, 3, 2, 0, 0, 1)^	N.A.	N.A.	N.A.	27.51 ±2.37^
	MCI [n=27]	68.04 ±7.32	22 (81.48%)	15.52 ±3.08	(24, 1, 1, 0, 0, 1)	N.A.	N.A.	N.A.	24.69 ±3.27^
	p-value	1.006e-2	1.115e-1	6.194e-1	2.910e-1	N.A.	N.A.	N.A.	3.004e-7
AIBL	NC [n=480]	72.45 ±6.22	203 (42.29%)	N.A.	N.A.	12 (2.50%)	28.70 ±1.24	0.03 ±0.12	N.A.
	MCI [n=102]	74.73 ±7.11	53 (51.96%)	N.A.	N.A.	12 (11.77%)	27.10 ±2.08	0.47 ±0.14	N.A.
	AD [n=79]	73.34 ±7.77	33 (41.77%)	N.A.	N.A.	14 (17.72%)	20.42 ±5.46	0.93 ±0.54	N.A.
	p-value	5.521e-3	1.887e-1	N.A.	N.A.	8.951e-9	4.585e-121	4.542e-158	N.A.
OASIS	NC [n=424]	71.34 ±9.43	164 (38.70%)	15.79 ±2.62^	(53, 18, 1, 0, 0, 0)^	121 (29.88%)	28.99 ±1.25^	0.00 ±0.02	N.A.
	MCI [n=27]	75.04 ±7.25	14 (51.85%)	15.19 ±2.76	(4, 1, 0, 0, 0, 0)^	9 (36.00%)	28.15 ±1.67	0.52 ±0.09	N.A.
	AD [n=193]	76.01 ±8.01	108 (55.96%)	14.68 ±3.09	(35, 9, 0, 0, 0, 0)^	102 (56.98%)	23.84 ±4.17	0.77 ±0.33	N.A.

	Non-AD [n=22]	72.64 ±8.77	16 (72.73%)	15.00 ±2.91	(6, 0, 0, 0, 0, 0)^	8 (47.06%)	24.14 ±4.69^	0.75 ±0.47	N.A.
	p-value	5.896e-8	3.190e-5	9.665e-5	8.098e-1	1.689e-9	2.122e-85	<1.0e-200	N.A.
FHS	NC [n=212]	73.37 ±9.63	112 (52.83%)	1.79 ±0.96 *	(207, 2, 1, 0, 0, 0)^	42 (20.19%) ^	28.14 ±1.72^	N.A.	N.A.
	MCI [n=75]	76.23 ±6.83	34 (45.33%)	1.59 ±0.98 *	(73, 0, 1, 0, 0, 0)^	17 (23.61%) ^	27.22 ±2.01^	N.A.	N.A.
	AD [n=17]	78.82 ±7.20	4 (23.53%)	1.82 ±0.92 *	(17, 0, 0, 0, 0, 0)	7 (43.75%) ^	24.00 ±2.13^	N.A.	N.A.
	Non-AD [n=9]	79.44 ±4.17	5 (55.56%)	1.00 ±1.15 *	(9, 0, 0, 0, 0, 0)	0 (0.00%)	22.00 ±2.45^	N.A.	N.A.
	p-value	4.755e-3	1.032e-1	5.918e-2	9.380e-1	5.704e-2	1.211e-13	N.A.	N.A.
LBDSU	NC [n=134]	68.77 ±7.62	61 (45.52%)	17.27 ±2.47^	N.A.	N.A.	N.A.	N.A.	27.43 ±2.23^
	MCI [n=35]	70.16 ±8.41	26 (74.29%)	16.60 ±2.58	N.A.	N.A.	N.A.	N.A.	24.00 ±3.14
	Non-AD [n=13]	73.42 ±7.81	8 (61.54%)	16.77 ±2.15	N.A.	N.A.	N.A.	N.A.	16.69 ±4.75
	p-value	1.033e-1	7.863e-3	3.243e-1	N.A.	N.A.	N.A.	N.A.	2.231e-30

Table 5.1. Study population and characteristics. Eight independent datasets were used for this study, including NACC, ADNI, AIBL, FHS, LBDSU, NIFD, OASIS, and PPMI. The NACC dataset was used to develop three separate types of models: i) an MRI-only CNN model that exclusively utilized imaging data, ii) non-imaging models in the form of traditional machine learning classifiers, which did not use any MRI data, and iii) a fusion model that combined imaging and non-imaging data within a hybrid architecture joining a CNN to a CatBoost model. The MRI-only model was validated across all eight cohorts, whereas external validation of non-imaging and fusion models was performed only on the OASIS cohort. All the MRI scans considered for this study were performed on individuals within ±6 months from the date of clinical diagnosis. The p-value for each dataset indicates the statistical significance of inter-group differences per column. We used two-tailed ANOVA and χ2 tests for continuous and categorical variables, respectively. NC = normal cognition, MCI = mild cognitive impairment, AD = Alzheimer's disease dementia, nADD = non-Alzheimer's disease dementia; NA = not available. *FHS education code: 0=high school did not graduate, 1=high school graduate, 2=some college graduate, 3=college graduate. The symbol ^ indicates that data was not available for some subjects.

We labeled the participants according to the clinical diagnosis. Subjects were labeled according to the clinical diagnoses provided by each study cohort (Appendix A. Diagnostic Criteria by Cohort). We kept MCI diagnoses without further consideration of underlying etiology to simulate a realistic spectrum of MCI presentations. For any subjects with documented dementia and primary diagnosis of Alzheimer's disease dementia, an AD label was assigned regardless of the presence of additional dementing comorbidities. Subjects with dementia but without confirmed ADD diagnosis were labeled as nADD. Notably, we elected to conglomerate all nADD subtypes into a singular label given that subdividing model training across an arbitrary number of prediction tasks ran the risk of diluting overall diagnostic accuracy. The ensemble of these 8 cohorts provided us a considerable number of participants with various forms of dementias as their primary diagnosis, including Alzheimer's disease dementia (ADD, n=1,606), Lewy body dementia (LBD, n=63), frontotemporal dementia (FTD, n=193), vascular dementia (VD, n=21), and other causes of dementia (n=237).

Subjects from each cohort were eligible for study inclusion if they had at least one T1-weighted volumetric MRI scan within 6 months of an officially documented diagnosis. We additionally excluded all MRI scans with fewer than 60 slices. For subjects with multiple MRIs and diagnosis records within a 6-month period, we selected the closest pairing of neuroimaging and diagnostic label. Therefore, only one MRI per subject was used. For the NACC and the OASIS cohorts, we further queried all available variables relating to demographics, past medical history, neuropsychological testing, and functional assessments. We did not use the availability of non-imaging features to exclude individuals in these cohorts and used K-nearest neighbor imputation for any missing data fields. Our overall data inclusion workflow may be found in Figure 5.2, where we reported the total number of subjects from each cohort before and after application of the inclusion criterion.

5.2.2 Harmonization of MRI Scans

To harmonize neuroimaging data between cohorts, we developed a pipeline of preprocessing operations (Figure 5.3) that was applied in identical fashion to all MRIs used in our study. This pipeline broadly consisted of two phases of registration to a standard MNI-152 template.

We describe Phase 1 as follows: (a) Scan axes were reconfigured to match the standard orientation of MNI-152 space. (b) Using an automated thresholding technique, a 3D volume-of-interest within the original MRI was identified containing only areas with brain tissue. (c) The volume-of-interest was skull-stripped to isolate brain pixels. (d) A preliminary linear registration of the skull-stripped brain to a standard MNI-152 template was performed. This step approximated a linear transformation matrix from the original MRI space to the MNI-152 space. Phase 2 was designed to fine-tune the quality of linear registration and parcellate the brain into discrete regions. These goals were accomplished by the following steps: (a) The transformation matrix computed from linear registration in Phase 1 was applied to the original MRI scan. (b) Skull stripping was once again performed after applying the linear registration computed from the initial volume of interest to isolate

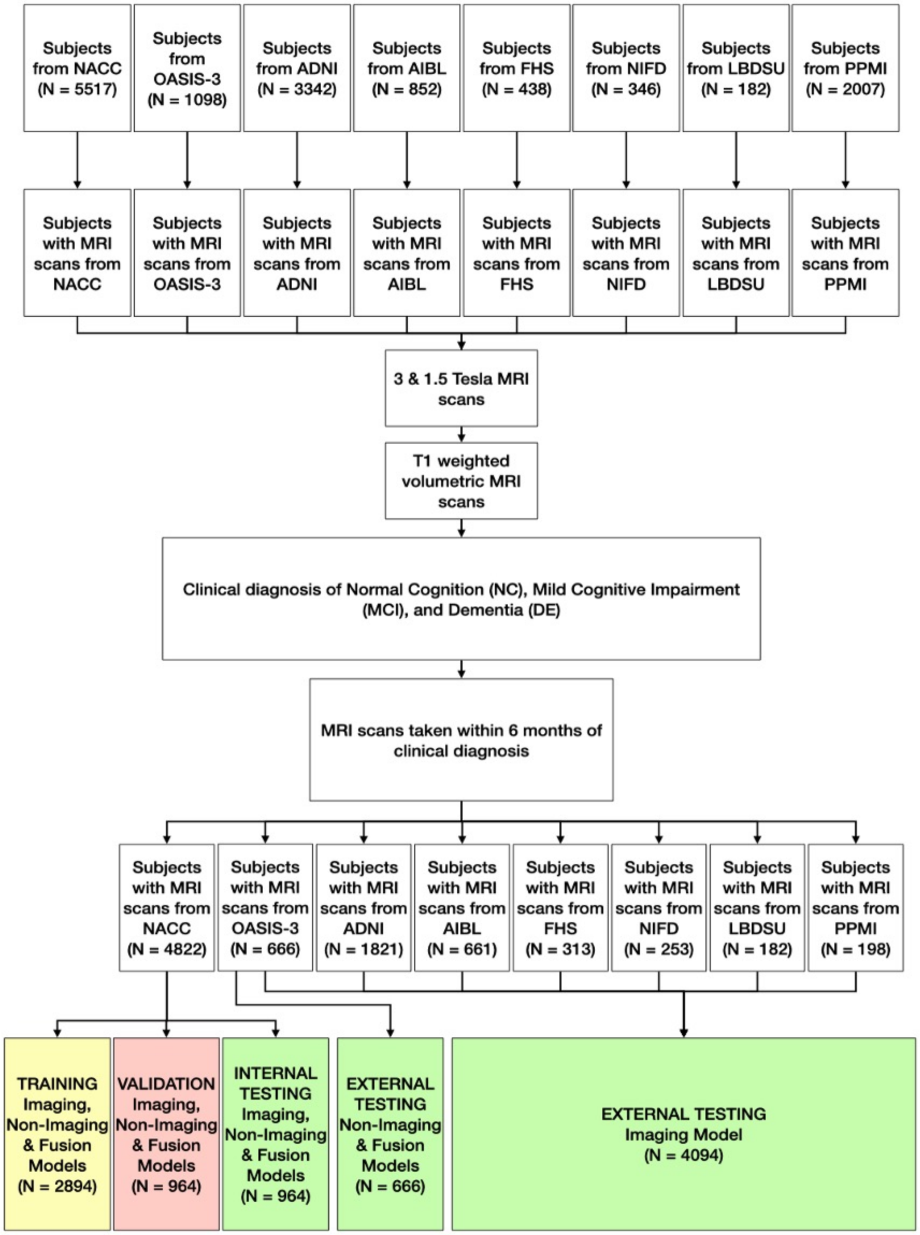

Subjects from NACC (N = 5517)
Subjects from OASIS-3 (N = 1098)
Subjects from ADNI (N = 3342)
Subjects from AIBL (N = 852)
Subjects from FHS (N = 438)
Subjects from NIFD (N = 346)
Subjects from LBDSU (N = 182)
Subjects from PPMI (N = 2007)
Subjects with MRI scans from NACC
Subjects with MRI scans from OASIS-3
Subjects with MRI scans from ADNI
Subjects with MRI scans from AIBL
Subjects with MRI scans from FHS
Subjects with MRI scans from NIFD
Subjects with MRI scans from LBDSU
Subjects with MRI scans from PPMI
3 & 1.5 Tesla MRI scans
T1 weighted volumetric MRI scans
Clinical diagnosis of Normal Cognition (NC), Mild Cognitive Impairment (MCI), and Dementia (DE)
MRI scans taken within 6 months of clinical diagnosis
Subjects with MRI scans from NACC (N = 4822)
Subjects with MRI scans from OASIS-3 (N = 666)
Subjects with MRI scans from ADNI (N = 1821)
Subjects with MRI scans from AIBL (N = 661)
Subjects with MRI scans from FHS (N = 313)
Subjects with MRI scans from NIFD (N = 253)
Subjects with MRI scans from LBDSU (N = 182)
Subjects with MRI scans from PPMI (N = 198)
TRAINING Imaging, Non-Imaging & Fusion Models (N = 2894)
VALIDATION Imaging, Non-Imaging & Fusion Models (N = 964)
INTERNAL TESTING Imaging, Non-Imaging & Fusion Models (N = 964)
EXTERNAL TESTING Non-Imaging & Fusion Models (N = 666)
EXTERNAL TESTING Imaging Model (N = 4094)

Figure 5.2. Data selection. Data from eight distinct study cohorts contributed towards the model development, validation, and testing: The NACC dataset (n=4,822), the OASIS-3 dataset (n=666), the ADNI dataset (n=1,821), the NIFD dataset (n=253), the PPMI dataset (n=198), the AIBL dataset (n=661), the FHS dataset (n=313), and in-house data from the LBDSU (n=182). In each dataset, T1-weighted, 1.5 and 3 Tesla MRIs were selected from participants. Only MRIs gathered within 6 months of MCI, AD or non-ADD diagnosis or last confirmed clinical visit (in the case of NC participants) were included for analysis.

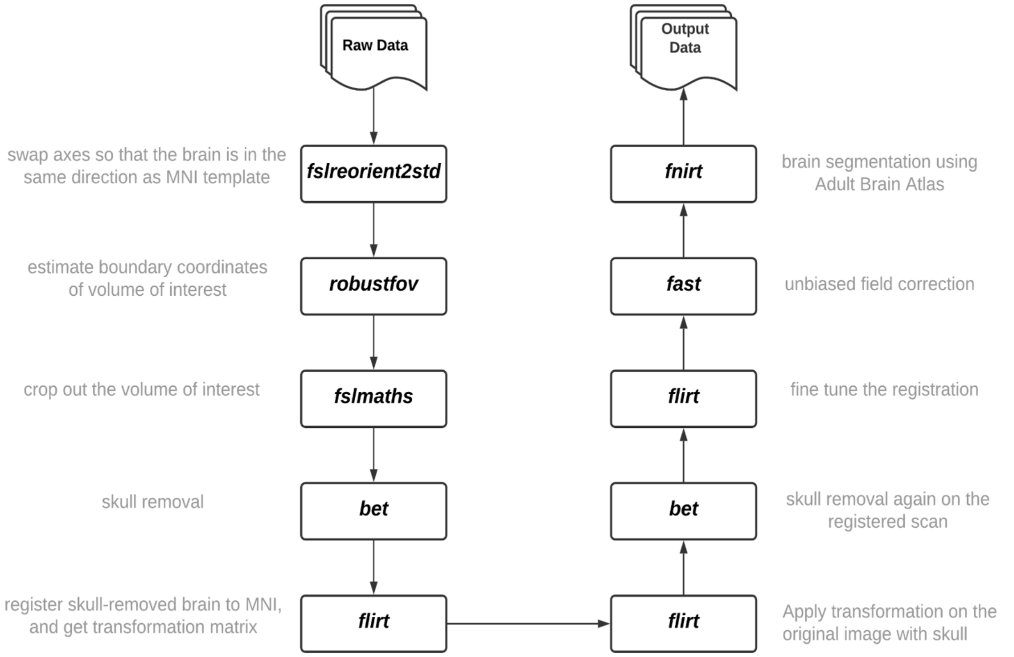

Raw Data
fslreorient2std
swap axes so that the brain is in the same direction as MNI template
robustfov
estimate boundary coordinates of volume of interest
fslmaths
crop out the volume of interest
bet
skull removal
flirt
register skull-removed brain to MNI, and get transformation matrix
flirt
Apply transformation on the original image with skull
bet
skull removal again on the registered scan
flirt
fine tune the registration
fast
unbiased field correction
fnirt
brain segmentation using Adult Brain Atlas
Output Data

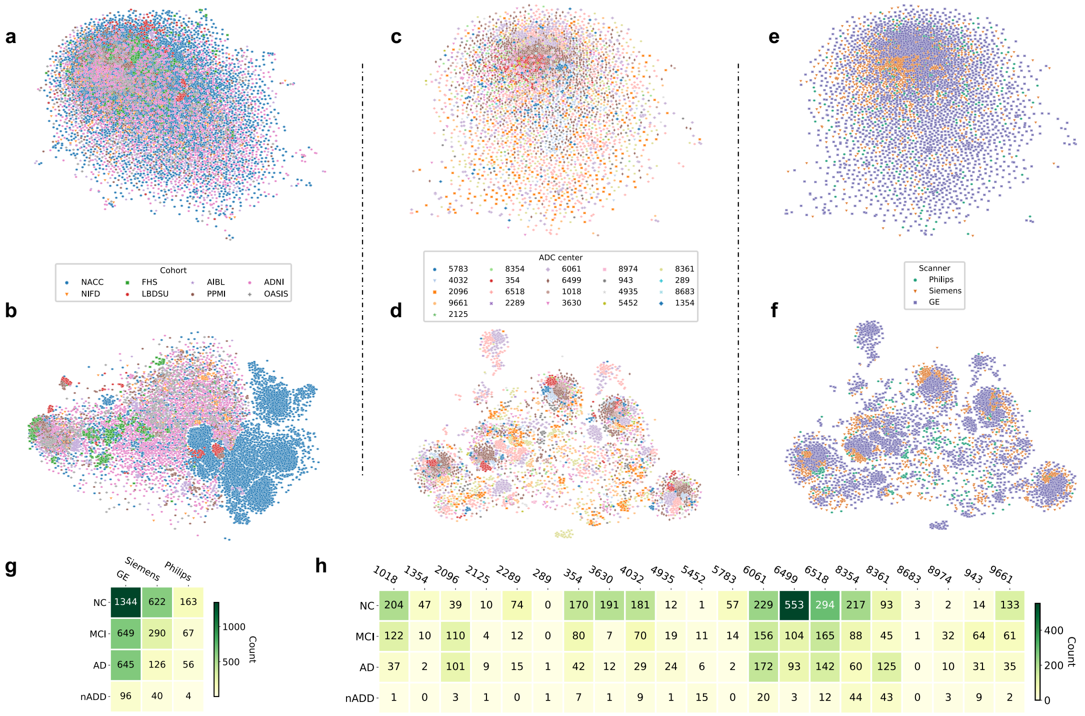
Figure 5.3. MRI preprocessing pipeline. MRI scans from all datasets were preprocessed using a common pipeline implemented in FSL. Raw MRIs were first reoriented to a standard axis layout and then aligned to the MNI-152 template using a linear registration tool and automatically identified region-of-interest. These aligned MRIs were then skull-stripped, and the resultant brains then underwent a second linear registration for fine-tuning of MNI alignments, as well as bias field correction for magnetic field inhomogeneities. Finally, specific brain regions were segmented by aligning the Hammersmith Adult brain atlas to registered brains using a non-linear registration. All processed MRIs were inspected visually, and individual brain extraction parameters were adjusted as needed for cases with failed registration. All FSL commands for the above steps are listed within boxes in the accompanying figure.

brain tissue from the full registered MRI scan. (c) Linear registration was applied again to alleviate any misalignments to MNI-152 space. (d) Bias field correction was applied to account for magnetic field inhomogeneities. (e) The brain was parcellated by applying a nonlinear warp of the Hammersmith Adult brain atlas to the postprocessed MRI.

All steps of our MRI-processing pipeline were conducted using FMRIB Software Library v6.0 (FSL) (Analysis Group, Oxford University). The overall preprocessing workflow was inspired by the harmonization protocols of the UK Biobank (https://git.fmrib.ox.ac.uk/falmagro/UK_biobank_pipeline_v_1). We manually inspected the outcome of the MRI pipeline on each scan to filter out cases with poor quality or significant processing artifacts.

5.2.3 Harmonization of Non-imaging Features

To harmonize the non-imaging variables across datasets, we first surveyed the available clinical data in all eight cohorts. We specifically examined information related to demographics, past medical history, neuropsychological test results, and functional assessments. Across a range of clinical features, we found the greatest availability of information in the NACC and the OASIS datasets. Additionally, given that the NACC and the OASIS cohorts follow Uniform Data Set (UDS) guidelines, we were able to make use of validated conversion scales between UDS versions 2.0 and 3.0 to align all cognitive measurements onto a common scale. We supply a full listing of clinical variables along with missing information rates per cohort in Appendix B. Feature Missing Rate.



	GE	Siemens	Philips
NC	1344	622	163
MCI	649	290	67
AD	645	126	56
nADD	96	40	4

	1018	1354	2096	2125	2289	289	354	3630	4032	4935	5452	5783	6061	6499	6518	8354	8361	8683	8974	943	9661
NC	204	47	39	10	74	0	170	191	181	12	1	57	229	553	294	217	93	3	2	14	133
MCI	122	10	110	4	12	0	80	7	70	19	11	14	156	104	165	88	45	1	32	64	61
AD	37	2	101	9	15	1	42	12	29	24	6	2	172	93	142	60	125	0	10	31	35
nADD	1	0	3	1	0	1	7	1	9	1	15	0	20	3	12	44	43	0	3	9	2

Figure 5.4. Site- and scanner-specific observations. (a) t-distributed stochastic neighbor embedding (tSNE) embeddings of down sampled MRI scans are shown. The down sampling was performed on the post-processed MRI scans using spline interpolation with a down sampling factor of 8 on each axis. (b) tSNEs of hidden-layer activations from the penultimate CNN hidden layer. Individual points correspond to internal representations of MRI scans during testing and are colored by cohort label. (c) tSNE of down sampled MRI scans from the NACC dataset is shown. Individual points are colored by the unique identifier of the Alzheimer Disease Research Centers (ADRCs) that participate in the NACC collaboration. (d) tSNE for penultimate-layer activations colored by ADRC ID are shown. (e) tSNE of down sampled MRI scans from the NACC dataset is shown. Color is assigned based on the manufacturer of the MRI scanner. (f) tSNE of penultimate layer activations is shown for cases in the NACC dataset. Embeddings are equivalent to those visualized in (d) but are now colored by the manufacturer of the scanner. (g) A tabular representation of disease category counts by manufacturer is presented. Only cases from the NACC dataset are included. We provide the Mutual Information Score (MIS) to quantify the correlation between disease type and scanner manufacturer. (h) We also provided a tabular representation of disease category counts stratified by ADRC ID in the NACC dataset. MIS is once again shown to quantify the degree of correlation between diagnostic labels and individual centers participating in the NACC study.

5.2.4 Effect of Confounding Factors

We further assessed our image harmonization pipeline by clustering the data using the t-distributed stochastic neighbor embedding (tSNE) algorithm[105]. We performed this procedure in order to ensure that (i) input data for all models was free of site-, scanner-, and cohort-specific biases and (ii) such biases could not be learned by a predictive model. To accomplish (i), we performed tSNE using pixel values from post-processed, 8x-downsampled MRI scans. For (ii), we performed tSNE using hidden-layer activations derived from the penultimate layer of a convolutional neural network (CNN) developed for our prediction tasks. For the NACC dataset, we assessed clustering of down sampled MRIs and hidden layer activations based on specific Alzheimer's Disease Research Centers (ADRCs) and scanner manufacturers (i.e., Siemens, Philips, and General Electric). We also repeated tSNE analysis based on specific cohorts (i.e., NACC, ADNI, FHS, etc.) using all available MRIs across our datasets. Using this approach, we observed no obvious clustering of postprocessed MRI embeddings among the eight cohorts used for testing of MRI-only models (Figure 5.4. a & b). Within the NACC cohort, we also observed no appreciable clustering based on individual Alzheimer's Disease Research Centers (ADRCs, Figure 5.4. c & d) or scanner manufacturer (Figure 5.4. e & f). Relatedly, although tSNE analysis of CNN hidden layer activations did yield clustering of NACC data points (Figure 5.4. b); this was an expected phenomenon given the selection of NACC as our cohort for model training. Otherwise, we appreciated no obvious conglomeration of embeddings from hidden layer activations due to specific ADRCs (Figure 5.4. d) or scanner manufacturers (Figure 5.4. f).

We also calculated mutual information scores (MIS) between ADRC ID, scanner brand, and diagnostic labels (NC, MCI, AD, and nADD) in the NACC dataset. This metric calculates the degree of similarity between two sets of labels on a common set of data. As with the tSNE analysis, the MIS calculation helped us to exclude the presence of confounding site- and scanner-specific biases on MRI data. Mutual Information Scores (MIS) computed from the NACC cohort indicated negligible correlation of diagnostic labels (NC, MCI, ADD, and nADD) between specific scanner manufacturers (MIS = 0.010, Figure 5.4. g) and ADRCs (MIS = 0.065, Figure 5.4. h).

5.3 Model development

5.3.1 Data Split and Model Architecture

We developed predictive models to meet two main objectives. The first, which we designated the COG task, was to predict the overall degree of cognitive impairment (either NC, MCI, or DE) in each participant based on neuroimaging. To meet this goal, we predicted a continuous 0-2 score (NC: 0, MCI: 1, DE: 2), which we denote as the DEmentia MOdel (DEMO) score. Of note, the COG task may also be regarded as consisting of three separate subtasks: (i) separation of NC from MCI and DE (COG_{NC} task), (ii) separation of MCI from NC and DE (COG_{MCI} task), and (iii) separation of DE from NC and MCI (COG_{DE} task). The second objective, which we designated the ADD task, was to predict whether a participant held a diagnosis of AD or nADD given that they were already predicted as DE in the COG task. For ease of reference, we denoted the probability of a person holding an AD diagnosis as the ALZheimer (ALZ) score. Following the sequential completion of the COG and ADD tasks, we were able to successfully separate ADD participants from NC, MCI, and nADD subjects.

We trained all models on the NACC dataset using cross validation. NACC was randomly divided into 5 folds of equal size with constant ratios of NC, MCI, AD, and nADD cases. We trained the model on 3 of the 5 folds and used the remaining two folds for validation and testing, respectively. Each tuned model was also tested on the full set of available cases from external datasets. Performance metrics for all models were reported as a mean across five folds of cross validation along with standard deviations and 95% confident intervals. Prior to training, we also set aside two specialized cohorts within NACC for neuropathologic validation and head-to-head comparison with clinicians. In the former case, we identified 74 subjects from whom post-mortem neuropathological data was available within 2 years of an MRI scan. In the latter, we randomly selected 100 age- and sex-matched groups of patients (25 per diagnostic category) to provide simulated cases to expert clinicians.

We used post-processed volumetric MRIs as inputs and trained a CNN model. To transfer information between the COG and ADD tasks, we trained a common set of convolutional blocks to act as general-purpose feature extractors. The DEMO and the ALZ scores were then calculated separately by appending respective fully connected layers to the shared convolutional backbone. We conducted the COG task as a regression problem using mean square error loss between the DEMO score and available cognitive labels. We performed the ADD task as a classification problem using binary cross entropy loss between the reference AD label and the ALZ score. The MRI-only model was trained using the NACC dataset and validated on all the other cohorts. To facilitate presentation of results, we pooled data from all the external cohorts (ADNI, AIBL, FHS, LBDSU, NIFD, OASIS and PPMI), and computed all the model performance metrics.

In addition to an MRI-only model, we developed a range of traditional machine learning classifiers using all available non-imaging variables shared between the NACC and the OASIS datasets. We first compiled vectors of demographics, past medical history, neuropsychological test results, and functional assessments. We scaled continuous variables by their mean and standard deviations and one-hot encoded categorical variables. These non-imaging data vectors were then passed as input to CatBoost, XGBoost, random forest, decision tree, multi-layer perceptron, support vector machine and K-nearest neighbor algorithms. Like the MRI-only model, each non-imaging model was sequentially trained to complete the COG and the ADD tasks by calculating the DEMO and the ALZ scores, respectively. We ultimately found that a CatBoost model yielded the best overall performance per area-under-receiver-operating-characteristic curve (AUC) and area-under-precision-recall curve (AP) metrics. We therefore selected this algorithm as the basis for follow-up analyses.

To mimic a clinical neurology setting, we developed a non-imaging model using data that is routinely collected for dementia diagnosis. A full listing of these variables used as input may be found in our Supplementary Information. While some features such as genetic status (APOE 4 allele), or cerebrospinal fluid measures have great predictive value, we have purposefully not included them for model development because they are not part of the standard clinical work-up of dementia.

To infer the extent to which completeness of non-imaging datasets influenced model performance, we conducted multiple experiments using different combinations of clinical data variables. The following combinations were input to the CatBoost algorithm for comparison: (1) demographic characteristics alone, (2) demographic characteristics and neuropsychological tests, (3) demographic characteristics and functional assessments, (4) demographic characteristics and past medical history, (5) demographic characteristics, neuropsychological tests and functional assessments, (6) demographic characteristics, neuropsychological tests and past medical history, and (7) demographic characteristics, neuropsychological tests, past medical history, and functional assessments.

To best leverage every aspect of the available data, we combined both MRI and non-imaging features into a common "fusion" model for the COG and the ADD tasks. The combination of data sources was accomplished by concatenating the DEMO and the ALZ scores derived from the MRI-only model to lists of clinical variables. The resultant vectors were then given as input to traditional machine learning classifiers as described above. Based on the AUC and the AP metrics, we ultimately found that a CNN linked with CatBoost model yielded the highest performance in discriminating different cognitive categories; the combination of CNN and CatBoost models was thus used as the final fusion model for all further experiments. Similarly, to our procedure with the non-imaging model, we studied how MRI features interacted with different subsets of demographic, past medical history, neuropsychological, and functional assessment variables. As with our non-imaging model, development and validation of fusion models was limited to NACC and OASIS only given limited availability of non-imaging data in other cohorts.

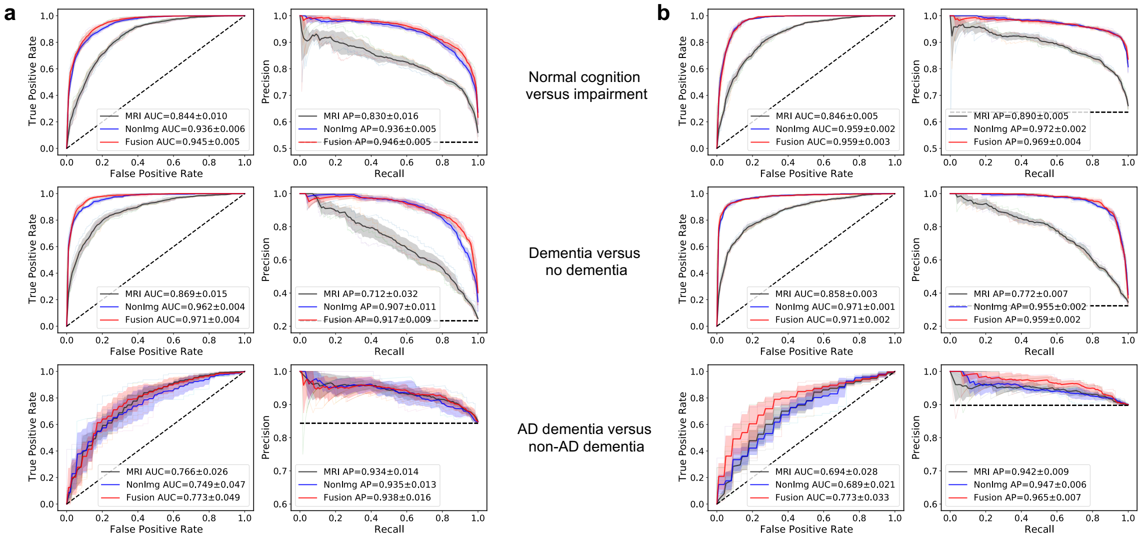


Figure 5.5. Performance of the deep learning models. (a-b) ROC curves showing true positive rate versus false positive rate and PR curves showing the positive predictive value versus sensitivity on the (a) NACC test set and (b) OASIS dataset. The first row in (a) and (b) denotes the performance of the MRI-only model, the non-imaging model, and the fusion model (CNN and CatBoost) trained to classify cases with NC from those without NC (COG$_{NC}$ task). The second row shows ROC and PR curves of the MRI-only model, the non-imaging model, and the fusion model for the COG$_{DE}$ task aimed at distinguishing cases with DE from those who do not have DE. The third row illustrates performance of the MRI-only model, the non-imaging model, and the fusion model focused on discriminating AD from nADD. For each curve, mean AUC was computed. In each plot, the mean ROC/PR curve and standard deviation are shown as bolded lines and shaded regions, respectively. The dotted lines in each plot indicate the classifier with the random performance level.

5.3.2 Model Performance

We observed that our fusion model provided the most accurate classification of cognitive status for NC, MCI, AD and nADD across a range of clinical diagnosis tasks. We found strong model performance on the COG_{NC} task between both the NACC test set (Figure 5.5. a, Row 1) and an external validation set (OASIS; Figure 5.5. b, Row 1) as indicated by area under the receiver operating characteristic (AUC) curve values of 0.945 [95% confidence interval (CI): 0.939, 0.951] and 0.959 [CI: 0.955, 0.963], respectively. Similar values for area under precision-recall (AP) curves were also observed, yielding 0.946 [CI: 0.940, 0.952] and 0.969 [CI: 0.964, 0.974], respectively. Such correspondence between AUC and AP performance supports robustness to class imbalance across datasets. In the COG_{DE} task, comparable results were also seen, as the fusion model yielded respective AUC and AP scores of 0.971 [CI: 0.966, 0.976]/0.917 [CI: 0.906, 0.928] (Figure 5.5. a, Row 2) in the NACC dataset and 0.971 [CI: 0.969, 0.973]/0.959 [CI: 0.957, 0.961] in the OASIS dataset (Figure 5.5. b, Row 2). Conversely, classification performance dropped slightly for the ADD task, with respective AUC/AP values of 0.773 [CI: 0.712, 0.834]/0.938 [CI: 0.918, 0.958] in the NACC dataset (Figure 5.5 a, Row 3) and 0.773 [CI: 0.732, 0.814]/0.965 [CI: 0.956, 0.974] in the OASIS dataset (Figure 5.5. b, Row 3).

Relative to the fusion model, we observed moderate performance reductions across classifications in our MRI-only model. For the COG_{NC} task, the MRI-only framework yielded AUC and AP scores of 0.844 [CI: 0.832, 0.856]/0.830 [CI: 0.810, 0.850] (NACC) and 0.846 [CI: 0.840, 0.852]/0.890 [CI: 0.884, 0.896] (OASIS). Model results were comparable on the COG_{DE} task, in which the MRI-only model achieved respective AUC and AP scores of 0.869 [CI: 0.850, 0.888]/0.712 [CI: 0.672, 0.752] (NACC) and 0.858 [CI: 0.854, 0.862]/0.772 [CI: 0.763, 0.781] (OASIS). For the ADD task as well, the results of the MRI-only model were approximately on par with those of the fusion model, giving respective AUC and AP scores of 0.766 [CI: 0.734, 0.798]/0.934 [CI: 0.917, 0.951] (NACC) and 0.694 [CI: 0.659,0.729]/0.942 [CI: 0.931, 0.953] (OASIS). For both fusion and MRI-only models, we also reported ROC and PR curves for the ADD task stratified by nADD subtypes in Appendix C. Classification Performance between AD and stratified nADD.

Interestingly, we note that a non-imaging model generally yielded similar results to those of both the fusion and MRI-only models. Specifically, a CatBoost model trained for the COG_{NC} task gave AUC and AP values 0.936 [CI: 0.929, 0.943] /0.936 [CI: 0.930, 0.942] (NACC), as well as 0.959 [CI: 0.957, 0.961]/0.972 [CI: 0.970, 0.974] (OASIS). Results remained strong for the COG_{DE} task, with AUC/PR pairs of 0.962 [CI: 0.957, 0.967]/0.907 [0.893, 0.921] (NACC) and 0.971 [CI: 0.970, 0.972]/0.955 [CI: 0.953, 0.957] (OASIS). For the ADD task, the non-imaging model resulted in respective AUC/PR scores of 0.749 [CI: 0.691, 0.807]/0.935 [CI: 0.919, 0.951] (NACC) and 0.689 [CI: 0.663, 0.715]/0.947 [CI: 0.940, 0.954] (OASIS). Performance of the MRI-only model across all external datasets is demonstrated via ROC and PR curves (Appendix D. MRI Model's ROC & PR Curves on 7 Cohorts).

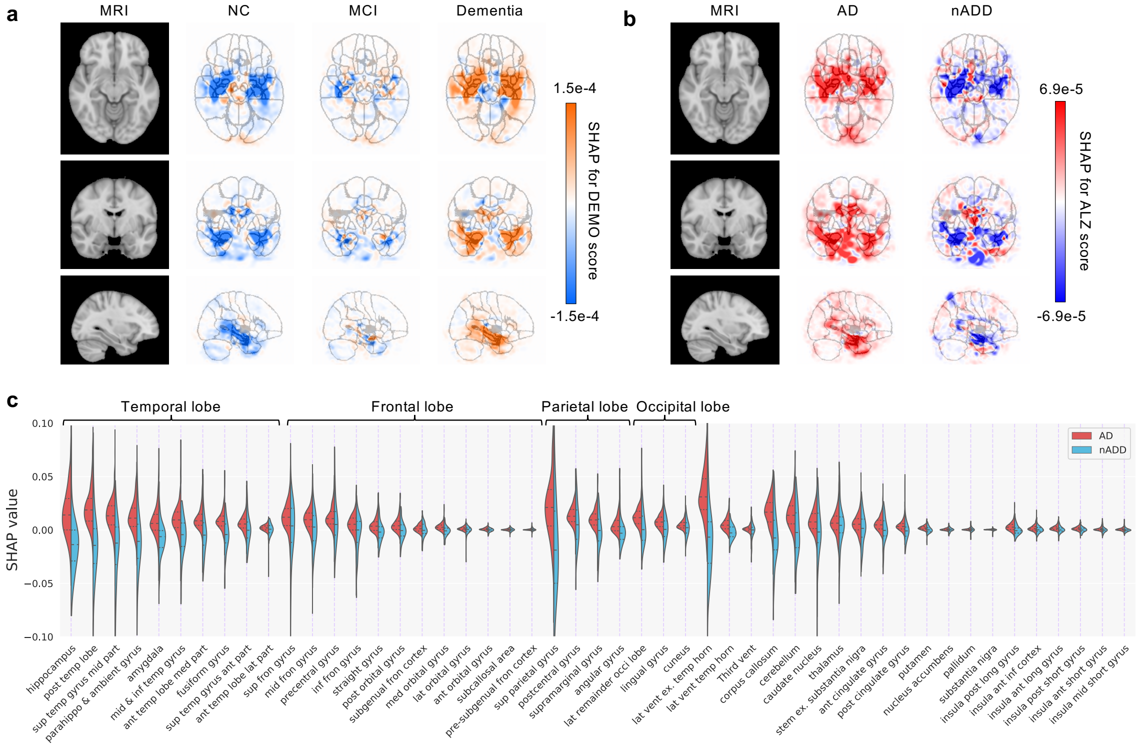


Figure 5.6. Neuroimaging signatures of dementia. (a-b) SHAP value-based illustration of brain regions that are most associated with the outcomes. The first columns in both (a) and (b) show a template MRI oriented in axial, coronal, and sagittal planes. In (a), the second, third and fourth columns show SHAP values from the input features of the second convolutional block of the CNN averaged across all NACC test subjects with NC, MCI, and dementia, respectively. In (b), the second and third columns show SHAP values averaged across all NACC test subjects with AD and nADD, respectively. (c) Brain region-specific SHAP values for both AD and nADD cases obtained from the NACC testing data are shown. The violin plots are organized per lobe and in decreasing order of mean absolute SHAP values.

5.4 SHAP-based Interpretability Analysis

In chapter 3, we introduced the SHAP method for interpreting ML models which was developed using the principle of classical Shapley value. The application of the SHAP-based interpretability analysis on both tabulate features and imaging features opens the opportunity of unveiling the underlying classification logic of the MRI model, the non-imaging model and the fusion model which were then compared with the expert impressions and ground truth biological evidence.

5.4.1 Voxel-level Saliency Map

The provenance of model predictions was visualized by pixel-wise SHAP mapping of hidden layers within the CNN model. Though a variety of methods exist for estimating SHAP values, we utilized a modified version of the DeepLIFT algorithm, which computes SHAP by estimating differences in model activations during backpropagation relative to a standard reference. We established this reference by integrating over a “background” of training MRIs to estimate a dataset-wide expected value. For each testing example, we then calculated SHAP values for the overall CNN model as well as for specific internal layers. Two sets of SHAP values were estimated for the COG and ADD tasks, respectively.

SHAP values calculated over the full model were directly mapped back to native MRI pixels whereas those derived for internal layers were translated to the native imaging space via nearest neighbor interpolation. The SHAP matrices were then correlated to physical locations within each subject’s MRI to visualize conspicuous brain regions implicated in each stage of cognitive decline from NC to dementia (Figure 5.6. a). This approach allowed neuroanatomical risk mapping to distinguish regions associated with AD from those with nADD (Figure 5.6. b). Indeed, the direct overlay of color maps representing disease risk on an anatomical atlas derived from traditional MRI scans facilitates interpretability of the deep learning model. Also, the uniqueness of the SHAP-derived representation allows us to observe disease suggestive regions that are specific to each outcome of interest.

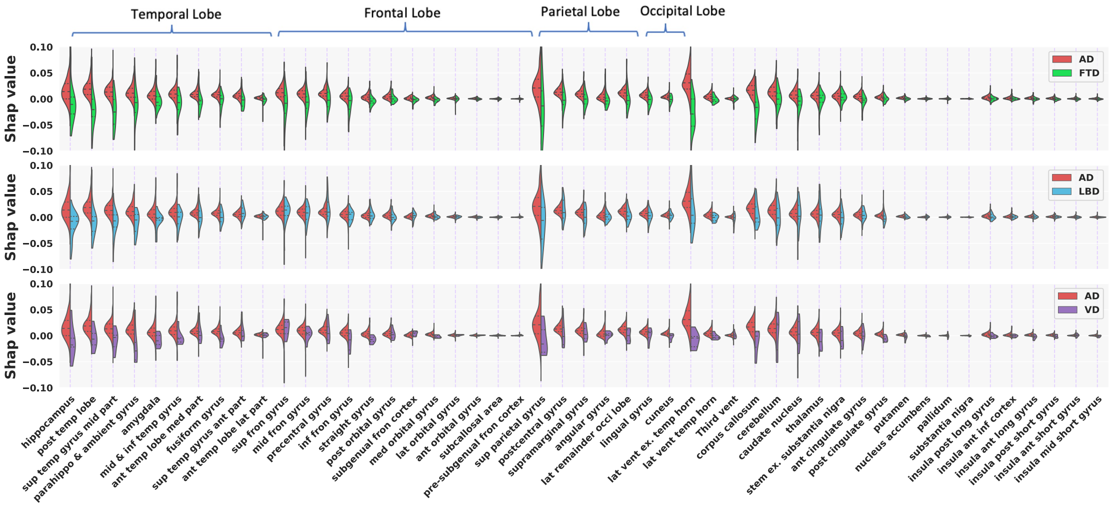


Figure 5.7. SHAP-based disease signature of stratified non-AD dementias. We presented the distribution of the regionally averaged SHAP values from the subjects that were correctly predicted as AD or non-AD dementias. The x-axis contains the region names that each violin plot is corresponding to. Comparisons on the regionally averaged SHAP distribution were made between AD and each non-AD dementias, including frontotemporal dementia (top row), Lewy body dementia (middle row), and vascular dementia (bottom row).

A key feature of SHAP is that a single voxel or a sub-region within the brain can contribute to accurate prediction of one or more class labels. For example, the SHAP values were negative in the hippocampal region in NC participants, but they were positive in participants with dementia, underscoring the well-recognized role of the hippocampus in memory function. Furthermore, positive SHAP values were observed within the hippocampal region for AD and negative SHAP values for the nADD cases, indicating that hippocampal atrophy has direct proportionality with AD-related etiology. The SHAP values sorted according to their importance on the parcellated brain regions also further confirm the role of hippocampus and its relationship with dementia prediction, particularly in the setting of AD (Figure 5.6. c), as well as nADD cases (Figure 5.7). In the case of nADD, the role of other brain regions such as the lateral ventricles and frontal lobes was also evident.

We also conducted a region-by-region graph analysis of SHAP values to determine whether consistent differences in ADD and nADD populations could be demonstrated (Figure 5.8). To visualize the relationship of SHAP scores across various brain regions, we created graphical representations of inter-region SHAP correlations within the brain. We derived region-specific scores by averaging voxel-wise SHAP values according to their location within the registered MRI. Subsequently, we constructed acyclic graphs in which nodes were defined as specific brain regions and edges as inter-regional correlations measured by Spearman's rank correlation and Pearson correlation coefficient, separately. To facilitate visualization and convey structural information, we manually aligned the nodes to a radiographic projection of the brain. The definition of all nodes can be found in Appendix E. Network Node Definition.

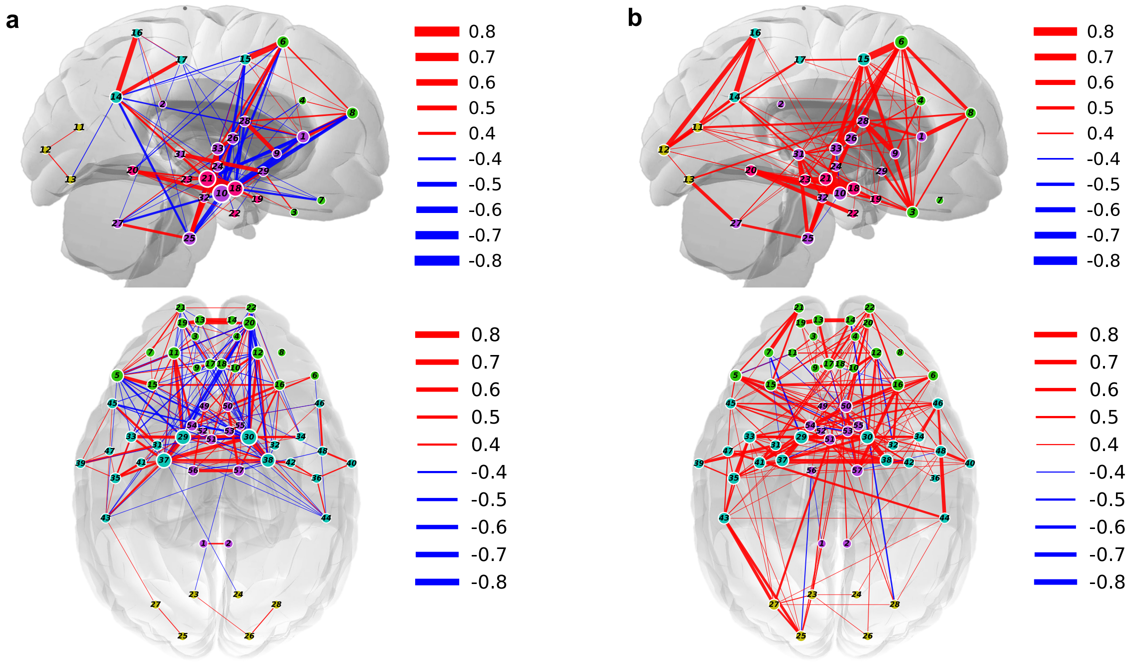


Figure 5.8. Network analysis of saliency maps over brain regions. The networks were generated using all (a) AD and (b) nADD subjects, respectively. We selected 33 representative brain regions for graph analysis and visualization of sagittal regions, as well as 57 regions for axial analyses. Nodes representing brain regions are overlaid on a two-dimensional brain template and sized according to weighted degree. The color of the segments connecting different nodes indicates the sign of correlation and the thickness of the segments indicates the magnitude of the correlation. It must be noted that not all nodes can be seen either from the sagittal or the axial planes.

Once correlation values were calculated between every pair of nodes, we filtered out the edges with p-value larger than 0.05 and ranked the remaining edges according to the absolute correlation value. We used only the top N edges (N=100 for sagittal view, N=200 for axial view) for the graph. We used color to indicate the sign of correlation and thickness to represent the magnitude of correlation. We used the following formula to derive the thickness:

$$thickness(corr.) = const. \times (abs(corr.) - threshold) \tag{11}$$

where the threshold is defined as the minimum of the absolute value of all selected edges' correlation value. The radius of nodes represents the weighted degree of the node which is defined as the sum of the edge weights for edges incident to that node. More specifically, we calculated the radius using the following equation:

$$radius(nodei) = 20 + 3 * (\sum_{j} correlation(node_i, node_j)) \tag{12}$$

In the above equation, we used 20 as a bias term to ensure that every node has at-least a minimal size to be visible on the graph. Note as well that the digit inside each node represents the index of the region name. Evidently, SHAP-based network analysis revealed pairwise relationships between brain regions that simultaneously contribute to patterns indicative of AD. The set of brain networks evinced by this analysis also demonstrate marked differences in structural changes between AD and nADD.

5.4.2 Feature Importance Ranking

To assess the contribution of various imaging and non-imaging features to classification outcomes, we calculated fifteen features with highest mean absolute SHAP values for the COG (Figure 5.9. a) and the ADD prediction tasks using the fusion model (Figure 5.9. b). Though MMSE score was the primary discriminative feature for the COG task, the DEMO score derived from the CNN portion of the model ranked third in predicting cognitive status. Analogously, the ALZ score derived from the CNN was the most salient feature in solving the ADD task.

Interestingly, the relative importance of features remained largely unchanged when a variety of other machine learning classifiers were substituted to the fusion model in lieu of the CatBoost model (Figure 5.9. c-d). This consistency indicated that our prediction framework was robust to the specific choice of model architecture, and instead relied on a consistent set of clinical features to achieve discrimination between NC, MCI, AD, and nADD classes. Relatedly, we also observed that non-imaging and fusion models retained predictive performance across a variety of input feature combinations, showing flexibility to operate across differences in information availability. Importantly, however, the addition of MRI-derived DEMO and ALZ scores improved 4-way classification performance across all combinations of non-imaging variables (Appendix F. Model Performance with Different Feature Combinations).

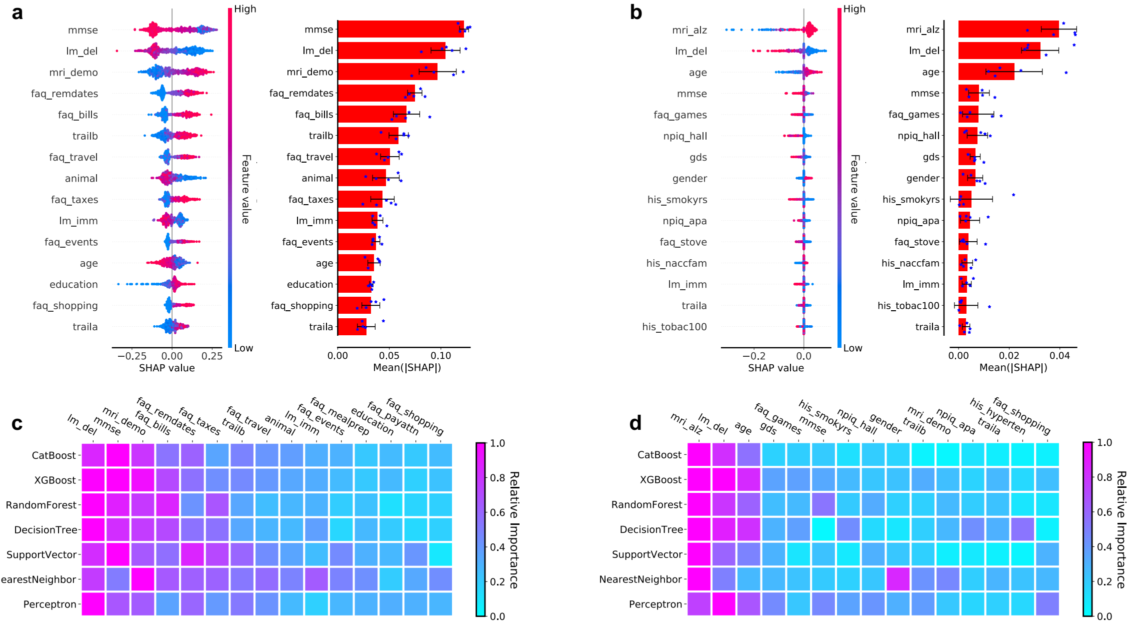


Figure 5.9. Feature importance. (a-b) Fifteen features with highest mean absolute SHAP values from the fusion model are shown for the COG and ADD tasks, respectively across cross-validation rounds (n=5). Error bars overlaid on bar plots are centered at the mean of the data and extend +/- one standard deviation. For each task, the MRI scans, demographic information, medical history, functional assessments, and neuropsychological test results were used as inputs to the deep learning model. The left plots in (a) and (b) illustrate the distribution of SHAP values and the right plots show the mean absolute SHAP values. All the plots in (a) and (b) are organized in decreasing order of mean absolute SHAP values. (c-d) For comparison, we also constructed traditional machine learning models to predict cognitive status and AD status using the same set of features used for the deep learning model, and the results are presented in (c) and (d), respectively. The heat maps show fifteen features with the highest mean absolute SHAP values obtained for each model.

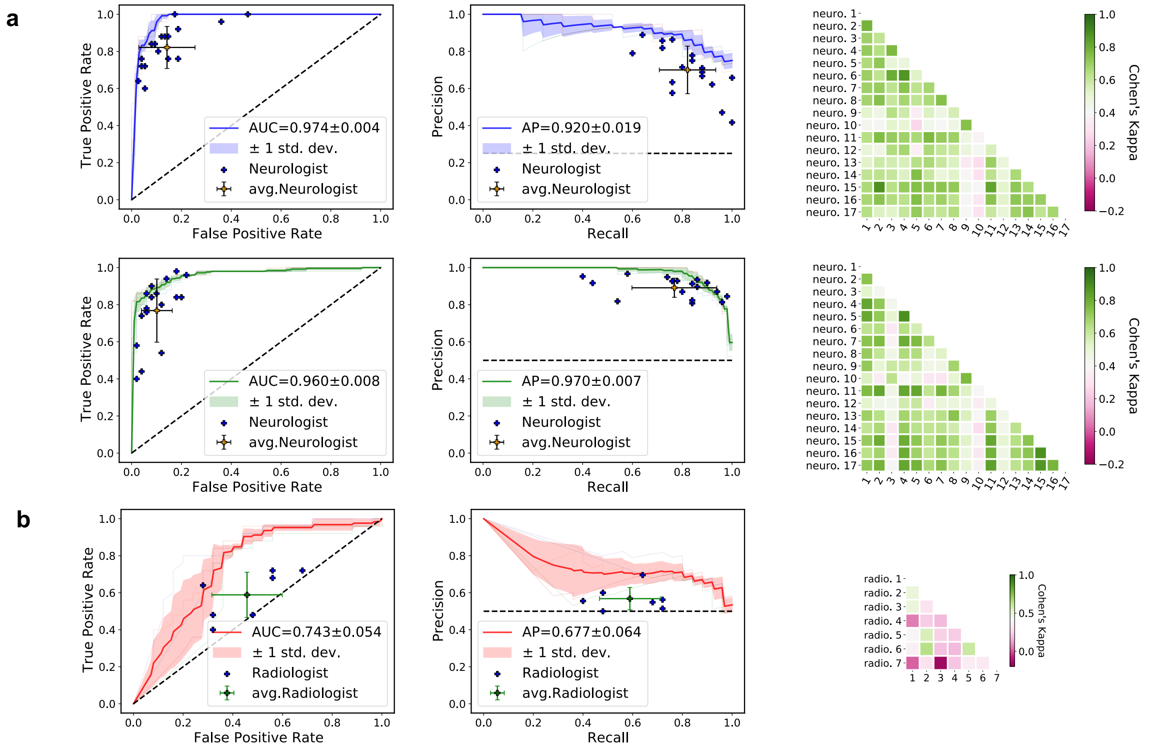


Figure 5.10. Expert-level validation. (a) For the COG_{NC} task (first row), the diagnostic accuracy of board-certified neurologists (n=17) is compared to the performance of our deep learning model using a random subset of cases from the NACC dataset (n=100). Metrics from individual clinicians are plotted in relation to the ROC and PR curves from the trained model. Individual clinician performance is indicated by the blue plus symbol and averaged clinician performance along with error bars is indicated by the green plus symbol on both the ROC and PR curves. The mean ROC/PR curve and the standard deviation are shown as the bold line and shaded region, respectively. A heatmap of pairwise Cohen's kappa statistic is also displayed to demonstrate inter-rater agreement across the clinician cohort. (a) For the COG_{DE} task (second row), ROC, PR, and interrater agreement graphics are illustrated with comparison to board-certified neurologists in identical fashion to (a). For both (a) and (b), all neurologists were granted access to multimodal patient data, including MRIs, demographics, medical history, functional assessments, and neuropsychological testing. The same data was used as input to train the deep learning model.

5.5 Validation

5.5.1 Expert-level Validation

To provide clinical benchmarking of our modeling approach, both neurologists and neuroradiologists were recruited to perform diagnostic tasks on a subset of NACC cases. The approach and performance of the neurologists and the neuroradiologists indicated variability across different clinical practices, with a moderate inter-rater agreement as evaluated using pairwise kappa (κ) scoring for all the tasks.

Among neurologists specifically, we observed average κ=0.600 for the COG_{NC} task (Figure 5.10. a, Row 1) and average κ=0.601 for the COG_{DE} task (Figure 5.10. a, Row 2). Among neuroradiologists performing the ADD task, we found average κ=0.292 (Figure 5.10. b). In the overall 4-way classification of NC, MCI, AD, and nADD, we observed that the accuracy of our fusion model (mean: 0.558, 95% CI: [0.582,0.634]) reached that of neurologists (mean: 0.565, 95% CI: [0.529,0.601]). Interestingly, a similar level of 4-way accuracy was achieved by a non-imaging CatBoost model (mean: 0.544, 95% CI: [0.517,0.571]), though not on an MRI-only model (mean: 0.412, 95% CI: [0.380,0.444]). However, an MRI-only model did yield a moderate improvement in diagnostic accuracy (mean: 0.692, 95% CI: [0.649,0.735]) over neuroradiologists (mean: 0.566, 95% CI: [0.516,0.616]) in the ADD task (Figure 5.10. b). Full performance metrics (including accuracy, sensitivity, specificity, F-1 score, and Matthews Correlation Coefficient) may be found in *Appendix G. Model vs Experts Performance* for respective comparison of machine learning models to neurologists and neuroradiologists in diagnostic simulations.

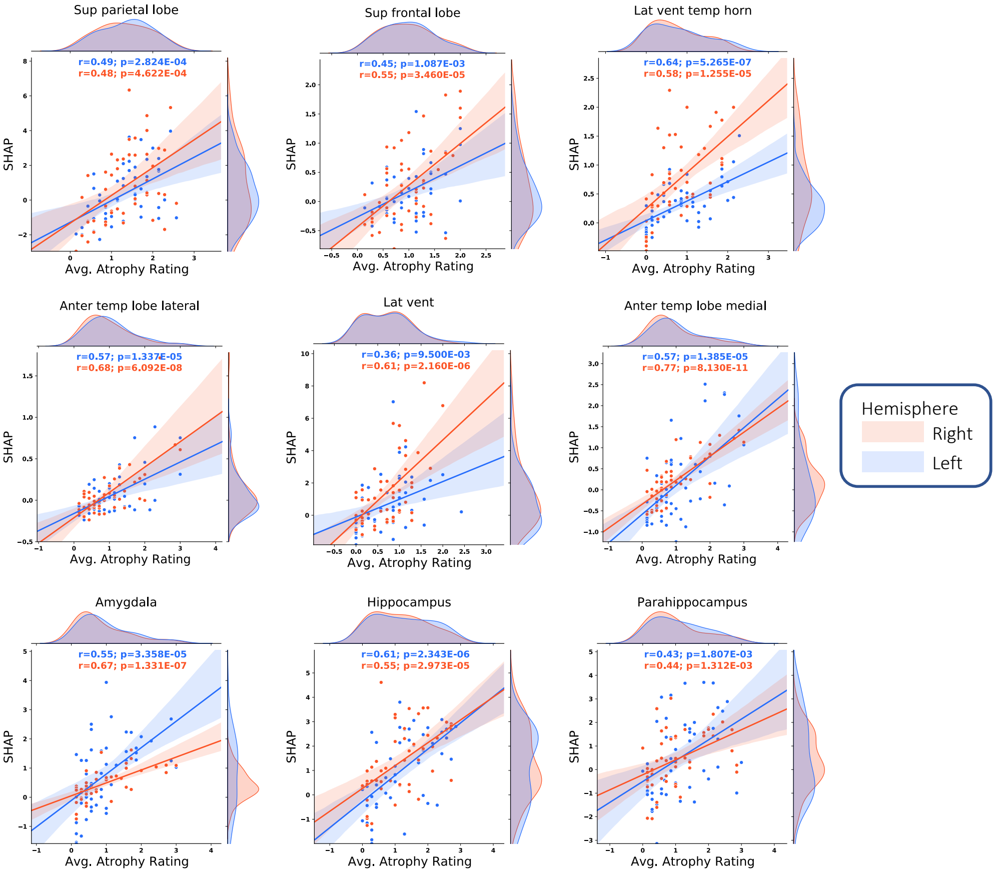


Figure 5.11. Validation against expert assessments on atrophy. SHAP values from the second convolutional layer averaged from selected brain regions are shown plotted against atrophy scores assigned by neuroradiologists. Orange and blue points (and along with regression lines and 95% confidence intervals) represent left and right hemispheres, respectively. Spearman correlation coefficients and corresponding two-tailed p-values are also shown and demonstrate a statistically significant proportionality between SHAP scores, and the severity of regional atrophy assigned by clinicians.

Performance metrics for simple thresholding of various neuropsychologic test scores can be found in Appendix H. Classification Performance from Simple Thresholding. We also sought to correlate region-specific SHAP values with structural changes observed by the neuroradiologists throughout the brain, with particular attention towards limbic and temporal lobe structures. Statistically significant correlations between regional SHAP averages and clinically graded atrophy severity suggested a connection between CNN features and widely known markers of dementia (Figure 5.11).

5.5.2 Neuropathological Validation

In addition to mapping hidden layer SHAP values to original neuroimaging, correlation of deep learning predictions with neuropathology data provided further validation of our modeling approach (Figure 5.12). Qualitatively, we observed that areas of high SHAP scores for the COG task correlated with region-specific neuropathological scores obtained from autopsy (Figure 5.12. a). Similarly, the severity of regional neuropathologic changes in these persons demonstrated a moderate to high degree of concordance with the regional cognitive risk scores derived from our CNN using the Spearman's rank correlation test. Of note, the strongest correlations appeared to occur within areas affected by AD pathology such as the temporal lobe, amygdala, hippocampus, and parahippocampal gyrus (Figure 5.12. b). Using the one-way ANOVA test, we also rejected a null hypothesis of there being no significant differences in DEMO scores between semi-quantitative neuropathological score groups (0-3) with a confidence level of 0.95, including for the global ABC severity scores of Thal phase for Aβ (A score F-test:

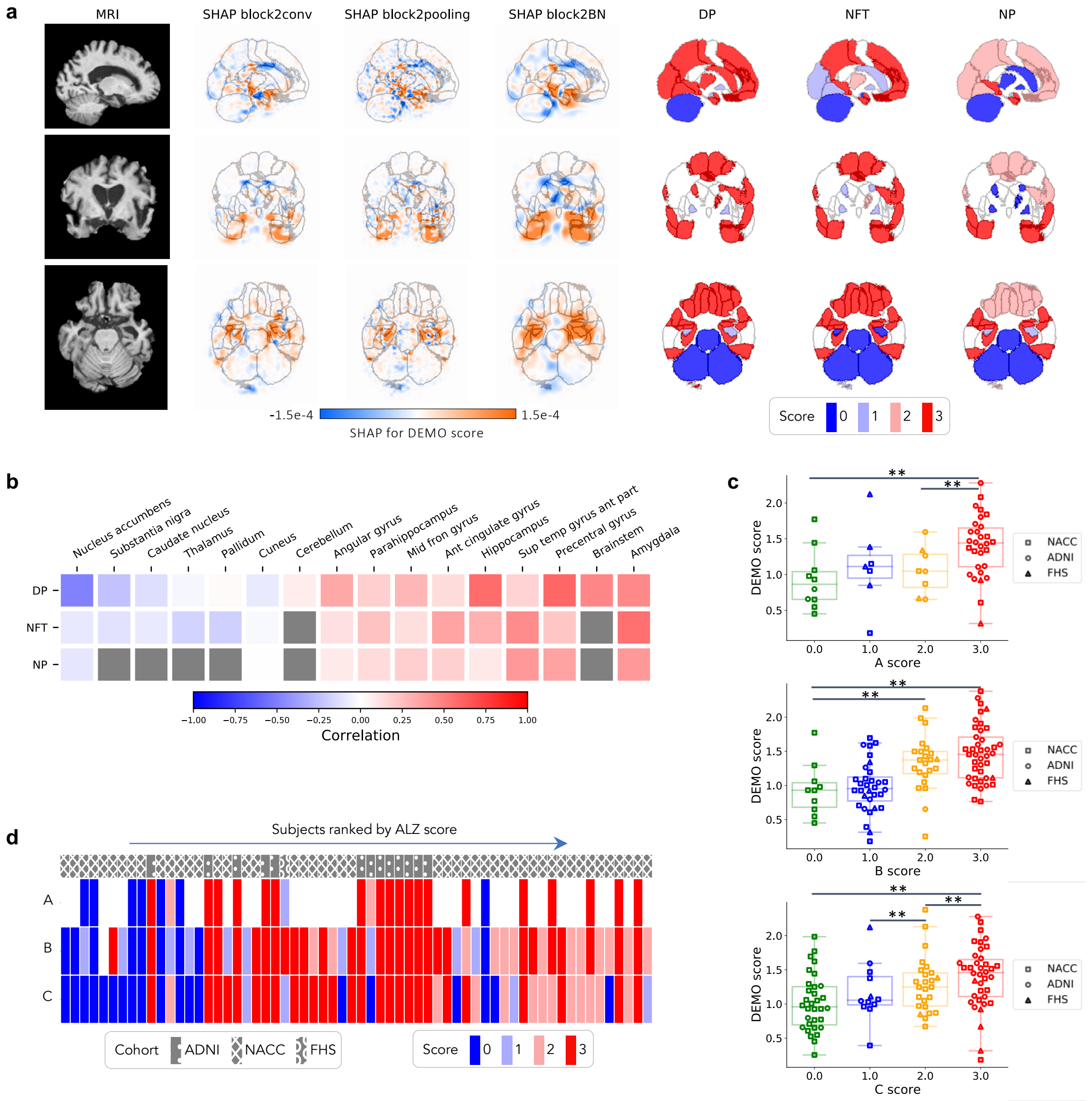


Figure 5.12. Neuropathological validation. (a) An example case from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset is displayed in sagittal, axial, and coronal views. The SHAP values derived from the second convolutional block and neuropathologic ABC scores are mapped to brain regions where they were measured at the time of autopsy. Visually, high concordance is observed between anatomically mapped SHAP values regardless of the hidden layer from which they are derived. Concordance is observed between the SHAP values and neurofibrillary tangles (NFT) scores within the temporal lobe. (b) A heatmap is shown demonstrating Spearman correlations between population-averaged SHAP values from the input features of the second convolutional layer and stain-specific ABC scores at various regions of the brain. A strong positive correlation is observed between the SHAP values and neuropathologic changes within several areas well-known to be affected in AD such as the hippocampus or parahippocampus, amygdala and temporal gyrus. (c) Beeswarm plots with overlying box-and-

whisker diagrams are shown to denote the distribution of ABC system sub-scores (horizontal axis) versus model-predicted cognitive scores (vertical axis). The displayed data points represent a pooled set of participants from ADNI, NACC, and FHS (n=118) for whom neuropathology reports were available from autopsy. Each symbol represents a study participant, boxes are centered at the median and extend over the interquartile range (IQR), while bottom and top whiskers represent 1st and 3rd quartiles -/+ 1.5 x IQR, respectively. Note: We denote $p < 0.05$ as *; $p<0.001$ as **, and $p < 0.0001$ as *** based on post-hoc Tukey testing. (d) A heatmap demonstrating the distribution of neuropathology scores versus model predicted AD probabilities. Herein, each column within the map represents a unique individual whose position along the horizontal axis is a descending function of AD risk according to the deep learning model. The overlying hatching pattern represents the dataset (ADNI, NACC and FHS), from which everyone is drawn.

pA=0.0014F(3, 51)=3.665, P-value=1.813e-2), Braak & Braak for neurofibrillary tangles (NFTs) (B score F-test: pB=0.0043F(3, 102)=11.528, P-value=1.432e-6), and CERAD neuritic plaque scores (C score F-test: pC=0.00010F(3, 103)=4.924, P-value=3.088e-3) (Figure 5.12. c). Of note, we also observed that the trend of ascending neuropathological score as the ALZ scores increases also correlated with semi-quantitative neuropathological scores (Figure 5.12. d).

5.6 Conclusion

In this chapter, we presented a range of machine learning models that can process multimodal clinical data to accurately perform a differential diagnosis of AD. These frameworks can achieve multiple diagnostic steps in succession, first delineating persons based on overall cognitive status (NC, MCI, and DE) and then separating likely cases of AD from those with nADD. Importantly, our models are capable of functioning with flexible combinations of imaging and non-imaging data, and their performance generalized well across multiple datasets featuring a diverse range of cognitive statuses and dementia subtypes.

The fusion model demonstrated the highest overall classification accuracy across diagnostic tasks, achieving results on par with neurologists recruited from multiple institutions to complete clinical simulations. Notably, similar levels of performance were observed both in the NACC testing set, and in the OASIS external validation set. The MRI-only model also surpassed the average diagnostic accuracy of practicing neuroradiologists and maintained a similar level of performance in 6 additional external cohorts (ADNI, AIBL, FHS, NIFD, PPMI, and LBDSU), thereby suggesting that diagnostic capability was not biased to any single data source. It is also worth noting that the DEMO and the ALZ scores bore strong analytic importance like that of traditional information used for dementia diagnosis. For instance, in the ADD task, the ALZ score was shown by SHAP analysis to have a greater impact in accurately predicting disease status than key demographic and neuropsychological test variables used in standard clinical practice such as age, sex, and MMSE score. These CNN-derived scores maintained equal levels of importance when used in other machine learning classifiers, suggesting wide utility for digital health workflows.

Furthermore, post-hoc analyses demonstrated that the performance of our machine learning models was grounded in well-established patterns of dementia-related neurodegeneration. Network analyses evinced differing regional distributions of SHAP values between AD and nADD populations, which were most pronounced in areas such as the hippocampus, amygdala, and temporal lobes. The SHAP values in these regions also exhibited a strong correlation with atrophy ratings from neuroradiologists. Although recent work has shown that explainable machine learning methods may identify spurious correlations in imaging data, we feel that our ability to link regional SHAP distributions to both anatomic atrophy and semi-quantitative scores of A-β amyloid, neurofibrillary tangles, and neuritic plaques links our modeling results to a gold-standard of postmortem diagnosis. More generally, our approach demonstrates a means by which to assimilate deep learning methodologies with validated clinical evidence in health care.

Our work builds on prior efforts to construct automated systems for the diagnosis of dementia. Previously, we developed and externally validated an interpretable deep learning approach to classify AD using multimodal inputs of MRI and clinical variables. Although this approach provided a novel framework, it relied on a contrived scenario of discriminating individuals into binary outcomes, which simplified the complexity of a real-world setting. Our current work extends this framework by mimicking a memory clinic setting and accounting for cases along the entire cognitive spectrum. Though numerous groups have taken on the challenge of nADD diagnosis using deep learning[109,110], even these tasks were constructed as simple binary classifications between disease subtypes. Given that the clinical practice of medicine rarely reduces to a choice between two pathologies, integrated models with the capability to replicate the differential diagnosis process of experts more fully are needed before deep learning models can be touted as assistive tools for clinical-decision support. Our results demonstrate a strategy for expanding the scope of diagnostic tasks using deep learning, while also ensuring that the predictions of automated systems remain grounded in established medical knowledge.

Interestingly, it should be noted that the performance of a non-imaging model alone approached that of the fusion model. However, the inclusion of neuroimaging data was critical to enable verification of our modeling results by clinical criteria (e.g., cross-correlation with post-mortem neuropathology reports). Such confirmatory data sources cannot be readily assimilated to non-imaging models, thus limiting the ability to independently ground their performance in non-computational standards. Therefore, rather than viewing the modest contribution of neuroimaging to diagnostic accuracy as a drawback, we argue that our results suggest a path towards balancing demands for transparency with the need to build models using routinely collected clinical data. Models such as ours may be validated in high-resource areas where the availability of advanced neuroimaging aids interpretability. As physicians may have difficulty entrusting medical decision-making to black box model in artificial intelligence[117], grounding our machine learning results in the established neuroscience of dementia may help to facilitate clinical uptake. Nevertheless, we note that our non-imaging model may be best suited for deployment among general practitioners (GPs) and in low-resource settings.

Functionally, we also contend that the flexibility of inputs afforded by our approach is a necessary precursor to clinical adoption at multiple stages of dementia. Given that subgroup analyses suggested significant 4-way diagnostic capacity on multiple combinations of training data (i.e., demographics, clinical variables, and neuropsychological tests), our overall framework is likely adaptable to many variations of clinical practice without requiring providers to significantly alter their typical workflows. For example, GPs frequently perform cognitive screening with or without directly ordering MRI tests[118–120], whereas memory specialists typically expand testing batteries to include imaging and advanced neuropsychological testing. This ability to integrate along the clinical care continuum, from primary to tertiary care allows our deep learning solution to address a two-tiered problem within integrated dementia care by providing a tool for both screening and downstream diagnosis.

This study has several limitations. To begin, in cases of mixed dementia, the present models default to a diagnosis of AD whenever this condition is present, thus attributing a single diagnosis to participants with multiple comorbidities. Given the considerable prevalence of mixed dementias[121], future work may include the possibility of a multi-label classification which may allow for the identification of co-occurring dementing conditions (e.g., LBD and AD, VD and AD) within the same individual. Our cohorts also did not contain any confirmed cases of atypical AD, which is estimated to affect approximately 6% of elderly-onset cases and one-third of patients with early-onset disease[122]. We must also note that MCI is a broad category by itself that includes persons who may or may not progress to dementia. When relevant data becomes available across many cohorts, future investigations could include MCI subjects who are amnestic and non-amnestic, to understand distinct signatures of those who have prodromal AD. We also acknowledge that our study data is predominantly obtained from epidemiologic studies which primarily focus on AD and that variables which optimize the identification of this illness may in fact detract from the accurate diagnosis of certain nADDs. For instance, we noted that the performance of our fusion models was slightly lower than that of the MRI-only model for distinguishing AD from non-parkinsonian dementias such as FTD and VD. We speculate that certain forms of neuropsychological testing such as the MMSE, which have well-known limitations in specificity[123], may bias predictions towards more common forms of dementia such as AD. Although we validated the various models with a population-based study (FHS) as well, it is possible that multimodal analysis frameworks have the potential to decrease diagnostic accuracy for less common dementias. Future modeling efforts may optimize for the identification of these diseases by including additional clinical data tailored to their diagnosis: for instance, the inclusion of motor examination to assess for parkinsonism, FLAIR images for vascular injury, or cognitive fluctuations and sleep behavior abnormalities for LBD. Lastly, although we have compared our model to the performance of individual neurologists and neuroradiologists, future studies may consider comparison to consensus reviews by teams of collaborating clinicians.

In conclusion, our interpretable, multimodal deep learning framework was able to obtain high accuracy signatures of dementia status from routinely collected clinical data, which was validated against data from independent cohorts, neuropathological findings, and expert-driven assessment. Moreover, our approach provides a solution that may be utilized across different practice types, from GPs to specialized memory clinics at tertiary care centers.

Chapter 6

Summary and Future Directions

In this dissertation, we developed and validated multiple deep learning frameworks that can process multimodal clinical data to accurately diagnose AD in different settings. With the long-term goal of creating an assistive clinical tool for dementia diagnosis, we primarily focused on establishing a systematic methodology that can be used as a guide to build a robust pathway towards the development of an AI-aided diagnostic tool. The work presented in this dissertation comprises three major components: (1) innovating technical methods in machine learning and computer vision that are tailored for the medical domain, (2) collaborating with medical professionals to assimilate their clinical acumen to model development, and (3) confirming model predictions with a more holistic view of the disease including the "gold standard" neuropathological findings and disease ratings from medical professionals.

We introduced both the traditional machine learning models and a variety of deep neural networks in chapter 2, where the basic concepts around training, validating, evaluating, and regularizing machine learning models were also discussed. Driven by improving accuracy of model predictions, the machine learning community is moving fast on innovating different aspects of the AI-based frameworks. As the predictive power increases, so does the complexity of such systems. Due to the black-box nature of these sophisticated deep neural networks, there remain obstacles to deploying machine learning systems in high-stakes scenarios. We witnessed a rapid increase of attention and interest in explainable and interpretable AI with the hope of elucidating the decision-making process of machine learning models. In chapter 3, we introduced a variety of interpretable machine learning methods and categorized these approaches based on different perspectives, i.e., model-agnostic or model specific, local, or global interpretation, as well as intrinsic or post-hoc interpretation. We also discussed the limitations of these natural image-based saliency methods when applied in healthcare domain, thus accentuating the need for medicine-specific computer vision methods.

In chapter 4, we presented a novel interpretable deep learning framework that overcomes some limitations of the CAM-based approaches and can be used to generate high-resolution saliency maps to delineate disease-specific signatures. The produced saliency maps are easy to interpret, since the saliency value directly represents the probability, as inferred from a patch of an MRI scan, that the subject has Alzheimer's disease. We coined the term disease probability map (DPM) to denote these representations. Different from the CAM-based saliency map whose value has arbitrary range, saliency values from DPM strictly range from 0 to 1, thus allowing convenient comparison of the disease severity across multiple instances. The statistically significant correlation between region-specific risk values from the DPM and region-specific biomarker measures from neuropathological examinations gave us confidence on the correspondence between the saliency outcomes with pathological changes in the brain. We also conducted a head-to-head comparison between the model and a group of practicing neurologists and showed that the model surpassed the average performance of 11 recruited neurologists in this specific binary classification task, i.e., identifying subjects with AD from those with normal cognition. Though we must acknowledge that this binary scenario is simpler than the real-world clinical diagnosis of AD, this work served as the first step towards building a more comprehensive diagnostic framework to characterize multiple etiologies of neurodegeneration. In addition, this work demonstrated the first case where three components of the proposed methodology, i.e., innovative technology, clinical acumen, and pathology confirmation, were developed, considered, and addressed in a single study, which uplifted our confidence in the validity of this deep learning framework.

In chapter 5, we expanded the diagnostic scope from the binary scenario, reported in chapter 4, to a more comprehensive classification between NC, MCI, AD, and nADD, which encompassed most individuals visiting memory clinics. We trained a multi-task deep learning model to predict (1) distinct levels of cognitive impairment and (2) various etiologies of dementia in two respective prediction tasks. This expanded diagnostic scope marked an important milestone towards the development of a clinical tool for differential diagnosis of AD, which requires care providers to consider all potential causes of relevant symptoms. To mimic the information availability in memory clinics, we used routinely collected clinical features, including demographic data, neuropsychological testing, functional assessments, medical history, and neuroimaging scans to construct our model. As the comprehensiveness of differential diagnosis increases, we might need to include, in the future projects, additional features tailored for the diagnosis of various brain diseases and disorders, e.g., emotional disorder, metabolic disorder, traumatic injury, infectious disease, depression, cerebrovascular disease, vitamin B12 deficiency etc. Because any of these diseases or disorders could co-exist in a patient and the clinical presentations of these conditions may also overlap, diagnosis is extremely challenging even for experts like neurologists. The future models, which could be potentially trained on a broader feature set and on multiple cohorts, might have enough diagnosing power to alleviate this burden.

Though medical professionals can order tests and scans if a critical piece of information is missing, diagnosis with partial information is still often done due to many reasons, such as the lack of resources to collect information, or the lack of expertise on interpreting novel measures. In chapter 5, we also reported multiple models trained with different combinations of feature sets to mimic distinct clinical scenarios, e.g., primary, and tertiary care settings. Special attention was paid to understanding the impact of including MRI-derived features to prediction accuracy, given multiple combinations of the non-imaging features were included. In the settings with limited non-imaging features, adding MRI-relevant information significantly boosted the model performance, whereas the effect started to diminish as more clinical features were included. In this dissertation, a KNN-based feature imputation was used to fill the missing feature values, which account for ~15% of all feature values, with the average of that feature values from K-nearest instances. However, there remain concerns around the quality of feature imputation. Future efforts should be made on exploring other options to handle missing features.

To further understand the relationship between input features and model predictions, we reported a series of analyses using a state-of-the-art interpretable machine learning method (SHAP) in chapter 5. With SHAP, we ranked the importance of all clinical features for different prediction tasks and assured the consistency of such results among seven distinct machine learning models. Instead of reapplying the same saliency method reported in chapter 4, we used SHAP-based saliency approach to “unbox” the convolutional neural network to highlight brain areas associated with dementia and AD. These SHAP-based saliency maps were then correlated with neuropathological findings and atrophy ratings from neuroradiologists to demonstrate the validity of the model predictions against pathological evidence and clinical acumen. We were also able to draw some data-driven insights using SHAP-based interpretability analysis, including identifying hippocampus as the most salient region for AD followed by overall identification of unique disease signatures of AD and nADD.

The model’s performance, as reported in chapter 6, was evaluated on multiple independent cohorts sourced from public datasets and in-house collaborations, including disease-specific and population-based studies. Consistent performance of our model on multiple independent cohorts demonstrated strong generalizability of the model predictions. To compare the model performance relative to that of dementia experts, we conducted head-to-head comparisons and demonstrated that the model was on par with experts on these pre-defined prediction tasks given the same amount of information. Nonetheless, we admit that the complexity of the designed prediction tasks is still not on par with that of the clinical scenario that neurologists are facing. To further increase the granularity of the differential diagnosis for AD, future efforts can be made to (1) subtype MCI into amnestic and non-amnestic categories, (2) disassemble nADD group into dementia etiologies other than AD, (3) allow mixed dementia diagnosis, and (4) include atypical dementia cases. Though we conducted a thorough search of data, the amount of data that normal research study can acquire is still not sufficient to develop an accurate model of differential diagnosis. Also, we hope to see more collaborative efforts throughout the research community on data collection, integration, and sharing to advance dementia research.

Because of the complexity of differential diagnosis of AD and dementia, we have not seen a single ML-based tool currently being used in memory clinics yet. Part of the reason is the lack of rigorous validation of such ML systems, which makes the use of ML models for final diagnosis unacceptable. Although the validation of ML frameworks using retrospective data can be done in multiple ways as reported in this dissertation, it is critical to test the model performance in a prospective setting to confirm if the model is as good as expected in real-world settings. Instead of aiming to use ML predictions to replace the diagnosis from clinicians, it is more realistic to think about how such systems can assist the doctors in different ways. Depending on the potential use cases, an ML model can be tuned towards having either a lower false negative rate (to accurately identify individuals with no signs of AD or dementia) or a lower false positive rate (to find those who likely have dementia or AD). If the model performance can be confirmed in both retrospective and prospective studies, then these ML frameworks can potentially be used as assistive tools to prioritize the subject queue so that the overall likelihood of providing timely diagnosis can be improved.

Appendix

Appendix A. Diagnostic Criteria by Cohort

Different dementia subtypes have different diagnostic guidelines and different centers may even have different criteria for inclusion in cohort studies. Given this variation, we herein provide summaries of such criteria for each of the datasets used for model development in our work. The information below is derived either directly from the documentation provided on each study's home webpage, or from highly cited articles that outline the requisite information.

ADNI

Subjects for ADNI were classified as NC, MCI, or mild AD. With respect to diagnostic criteria, NC subjects had no memory complaints while MCI and AD subjects both had to have had complaints. Participants undergo a full review of past medical history, medications, and provide blood for APOE DNA testing, and undergo a battery of neuropsychological tests at baseline. All subjects also undergo 1.5T MRI, with additional subsets selected for 3T MRI, PET, and lumbar puncture. Criteria for clinical classification of NC MCI and AD are as follows:

- **NC:**

  No memory complaints, aside from those common to other normal subjects of that age range. Normal memory function was documented by scoring at specific cutoffs on the Logical Memory II subscale (delayed Paragraph Recall) from the Wechsler Memory Scaled - Revised (the maximum score is 25):

a) greater than or equal to 9 for 16 or more years of education

b) greater than or equal to 5 for 8-15 years of education

c) greater than or equal to 3 for 0-7 years of education.

Mini-Mental State Exam score between 24 and 30 (inclusive) (Exceptions may be made for subjects with less than 8 years of education at the discretion of the project director). Clinical Dementia Rating = 0. Memory Box score must be 0. Cognitively normal, based on an absence of significant impairment in cognitive functions or activities of daily living

- **MCI**:

Memory complaint reported by subject or study partner that is verified by a study partner. Abnormal memory function was documented by scoring below the education adjusted cutoff on the Logical Memory II subscale (Delayed Paragraph Recall) from the Wechsler Memory Scale – Revised (the maximum score is 25):

a) less than or equal to 8 for 16 or more years of education

b) less than or equal to 4 for 8-15 years of education

c) less than or equal to 2 for 0-7 years of education.

Mini-Mental State Exam score between 24 and 30 (inclusive) (Exceptions may be made for subjects with less than 8 years of education at the discretion of the project director). Clinical Dementia Rating = 0.5. Memory Box score must be at least 0.5. General cognition and functional performance sufficiently preserved such that a diagnosis of Alzheimer’s disease cannot be made by the site physician at the time of the screening visit. MMSE between 24-30, memory complaint, objective memory loss measured by education-adjusted score on Wechsler Memory Scale Logical Memory II, CDR 0.5, preserved activities of daily living, absent dementia, and absence of significant cognitive impairment in other domains.

- **AD**:

  Memory complaint reported by subject or study partner that is verified by a study partner. Abnormal memory function was documented by scoring below the education adjusted cutoff on the Logical Memory II subscale (Delayed Paragraph Recall) from the Wechsler Memory Scale – Revised (the maximum score is 25):

  a) less than or equal to 8 for 16 or more years of education

  b) less than or equal to 4 for 8-15 years of education

  c) less than or equal to 2 for 0-7 years of education.

  MMSE between 20 and 26 (inclusive) (Exceptions may be made for subjects with less than 8 years of education at the discretion of the protocol PI). Clinical Dementia Rating = 0.5, 1.0. NINCDS/ADRDA criteria for probable AD

NACC

NACC employs a longitudinal data collection protocol using prospective, standardized clinical evaluation of subjects in the National Institute of Aging's Alzheimer's Disease Research Centers (ADRCs). Each center enrolls participants according to its own protocol. Subjects, along with their family and friends, if necessary, participate in annual screening questionnaires comprising NACC's Uniform Data Set (UDS) standards. Questionnaire results include neuropsychological testing, medical history, and present symptoms. Final diagnosis is made by either a consensus team of experts, or the examining physician, with the exact arbiter varying according to the specific ADRC.

PPMI

PPMI subjects are recruited at disease threshold, within two years of the time of first diagnosis by a treating clinician. While diagnostic criteria per se are not set forth in the study's documentation, participation for patients with Parkinson's disease (PD) is standardized by certain inclusion criteria, which include:

- At least two of them: bradykinesia, resting tremor, and rigidity OR either asymmetric resting tremor OR asymmetric bradykinesia.
- No prior treatment for PD
- Note expected to require PD medication within at least 6 months for baseline
- At least 30 years old at the time of PD diagnosis
- Hoehn and Yahr stage I or II at baseline
- Dopamine transporter deficit per SPECT or VMAT deficit per VMAT-2 PET scan
- Asymmetric resting tremor or asymmetric bradykinesia

LBDSU

Diagnosis of Lewy Body dementia from the Lewy Body Dementia Research Center of Excellence at Stanford University follows diagnostic criteria set forth by the consensus of the Dementia with Lewy Bodies Consortium. While a full accounting of Core Clinical Features, Supportive Clinical Features, Indicative Biomarkers, and Supportive Biomarkers may be found in this review paper[124], these guidelines define criteria for diagnosis of both probable and possible cases of Lewy Body dementia. In the present study, we utilized information from subjects meeting criteria for probably Lewy Body dementia, which is defined as either a) ≥ 2 core clinical features of DLB present with/without presence of indicative of biomarkers or b) only one core clinical feature present but with one or more indicative biomarkers[124].

AIBL

All volunteers in the AIBL study underwent a screening interview, comprehensive cognitive testing, health, and lifestyle questionnaires. Allocation of individuals to one of three diagnostic groups (NC, MCI, AD) was performed by a clinical review panel's consensus, which assessed both patients with a known history of MCI or AD diagnosis, as well as those recruited as NC who demonstrated any of the following conditions:

- MMSE < 28/30
- Failure on Logical Memory test
- Clinical Dementia Rating (CDR) score ≥ 0.5
- Medical history suggestive of the presence of illness is likely to impair cognitive function
- Informant or personal history suggestive of cognitive impairment
- Consumption of medications or other substances that could affect cognition
- Other evidence of significant cognitive difficulty on neuropsychological testing both

The clinical review panel consisted of two geriatric psychiatrists, a neurologist, a geriatrician, and five neuropsychologists. AD diagnoses included DSM-IV diagnostic criteria, ICD-10 dementia severity rating, and NINCDS-ADRDA diagnostic criteria. MCI diagnoses were made according to criteria set forth[125], and those presenting with previously diagnosed MCI were required to demonstrate a score of 1.5 standard deviations or more below age-adjusted mean on at least one neuropsychological test. Those reporting as NC were eligible for a diagnosis of MCI if they demonstrated performance at least 1.5 standard deviations on two or more neuropsychological tests in addition to having reported memory difficulties.

OASIS

The OASIS dataset includes participants enrolled into several ongoing studies through the Charles F. and Joanne Knight ADRC at Washington University in St. Louis. As in NACC, participants completed clinical assessments according to the UDS. Using the UDS, dementia status was assessed using the CDR score, with 0 indicating NC, 0.5 very mild impairment, 1 MCI, and 2 moderate dementia. Diagnostic impressions are also provided as a separate variable within this dataset by examining physicians and are available in separate data fields from the UDS-derived score.

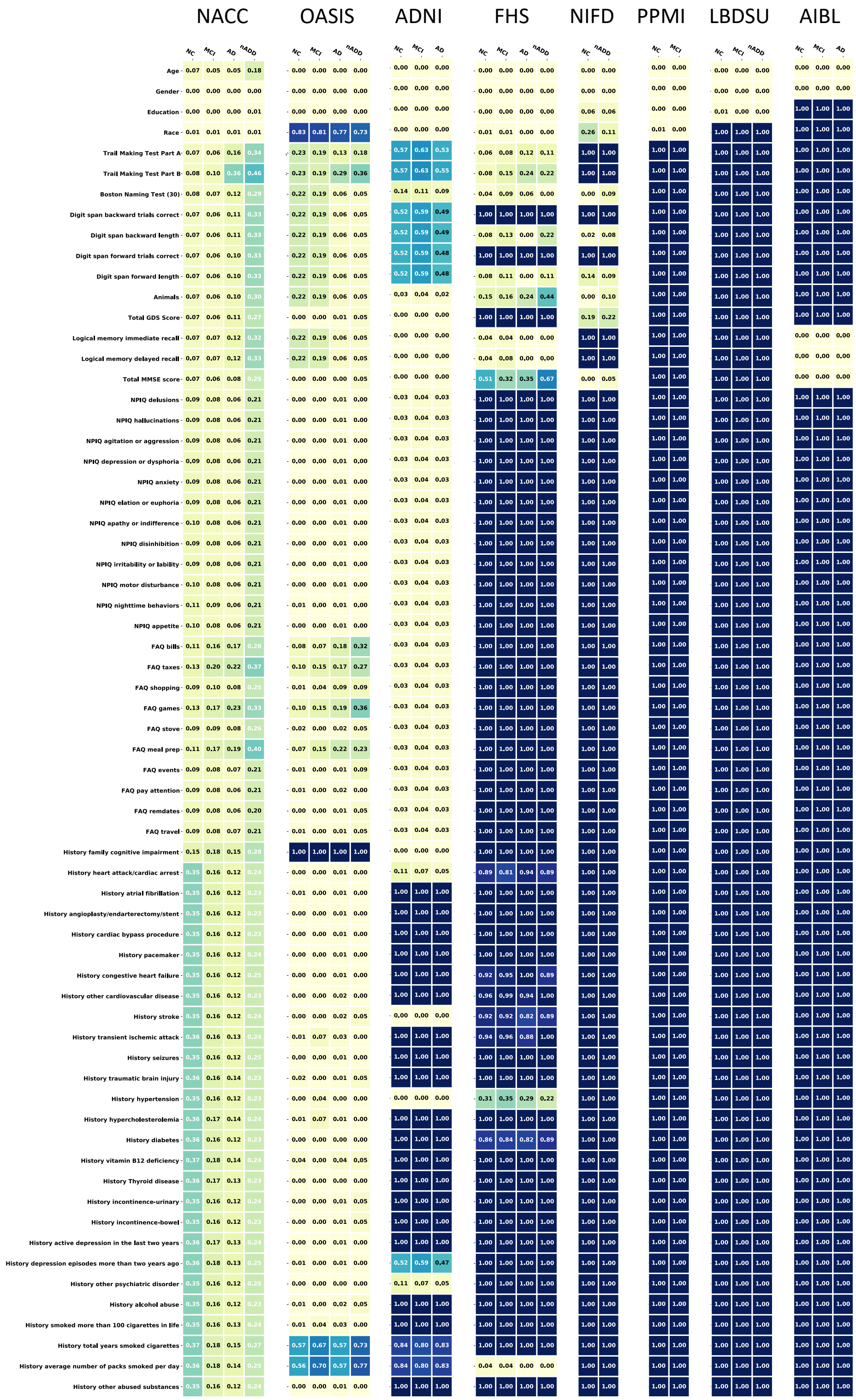
FHS

The original cohort of FHS participants underwent the Kaplan-Albert neuropsychological test battery in their fourteenth examination cycle and has since been monitored at regular intervals. Persons scoring below education-based cutoffs on MMSE or those who experience a decrease of at least three points between examinations are flagged for additional rounds of neurological and neuropsychological assessment. Family- or self-reported memory loss symptoms, as well as referrals from FHS physicians and staff may similarly lead participants to in-depth cognitive assessments. All follow-up testing that remains concerning for dementia is assessed by a panel consisting of at least one neurologist and one neuropsychologist, who utilize available neurological and neuropsychological testing, a structured telephone interview with family members or caregivers, past FHS and medical history, as well as imaging and autopsy results where available. Dementia and MCI are diagnosed according to DSM criteria, and AD specifically is diagnosed according to criteria from the National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders Association[126].

NIFD

The Frontotemporal Lobe Dementia Neuroimaging Initiative encompasses individuals diagnosed with behavioral variant Frontotemporal Dementia (bvFTD), semantic variant Primary Progressive Aphasia (svPPA), and non-fluent variant Primary Progressive Aphasia, as well as age-matched controls for each of those cohorts. Recruited patients are drawn from clinical sites at the University of California, San Francisco, Massachusetts General Hospital, and the Mayo Clinic of Minnesota. bvFTD diagnostic criteria are set forth in the recommendations of the International Behavioral Variant FTD Criteria Consortium[127], and criteria for diagnosis of primary progressive aphasia are similarly based upon international expert consensus which encompass clinical, imaging, and biomarker data[128].

Appendix B. Feature Missing Rate



	NACC				OASIS				ADNI			FHS				NIFD		PPMI		LBDSU			AIBL		
	NC	MCI	AD	nADD	NC	MCI	AD	nADD	NC	MCI	AD	NC	MCI	AD	nADD	NC	nADD	NC	MCI	NC	MCI	nADD	NC	MCI	AD
Age	0.07	0.05	0.05	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gender	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Education	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.06	0.00	0.00	0.01	0.00	0.00	1.00	1.00	1.00
Race	0.01	0.01	0.01	0.01	0.83	0.81	0.77	0.73	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.26	0.11	0.01	0.00	1.00	1.00	1.00	1.00	1.00	1.00
Trail Making Test Part A	0.07	0.06	0.16	0.34	0.23	0.19	0.13	0.18	0.57	0.63	0.53	0.06	0.08	0.12	0.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Trail Making Test Part B	0.08	0.10	0.36	0.46	0.23	0.19	0.29	0.36	0.57	0.63	0.55	0.08	0.15	0.24	0.22	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Boston Naming Test (30)	0.08	0.07	0.12	0.29	0.22	0.19	0.06	0.05	0.14	0.11	0.09	0.04	0.09	0.06	0.00	0.00	0.09	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Digit span backward trials correct	0.07	0.06	0.11	0.33	0.22	0.19	0.06	0.05	0.52	0.59	0.49	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Digit span backward length	0.07	0.06	0.11	0.33	0.22	0.19	0.06	0.05	0.52	0.59	0.49	0.08	0.13	0.00	0.22	0.02	0.08	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Digit span forward trials correct	0.07	0.06	0.10	0.33	0.22	0.19	0.06	0.05	0.52	0.59	0.48	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Digit span forward length	0.07	0.06	0.10	0.33	0.22	0.19	0.06	0.05	0.52	0.59	0.48	0.08	0.11	0.00	0.11	0.14	0.09	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Animals	0.07	0.06	0.10	0.30	0.22	0.19	0.06	0.05	0.03	0.04	0.02	0.15	0.16	0.24	0.44	0.00	0.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Total GDS Score	0.07	0.06	0.11	0.27	0.00	0.00	0.01	0.05	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.19	0.22	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Logical memory immediate recall	0.07	0.07	0.12	0.32	0.22	0.19	0.06	0.05	0.00	0.00	0.00	0.04	0.04	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00
Logical memory delayed recall	0.07	0.07	0.12	0.33	0.22	0.19	0.06	0.05	0.00	0.00	0.00	0.04	0.08	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00
Total MMSE score	0.07	0.06	0.08	0.25	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.51	0.32	0.35	0.67	0.00	0.05	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00
NPIQ delusions	0.09	0.08	0.06	0.21	0.00	0.00	0.01	0.00	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NPIQ hallucinations	0.09	0.08	0.06	0.21	0.00	0.00	0.01	0.00	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NPIQ agitation or aggression	0.09	0.08	0.06	0.21	0.00	0.00	0.01	0.00	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NPIQ depression or dysphoria	0.09	0.08	0.06	0.21	0.00	0.00	0.01	0.00	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NPIQ anxiety	0.09	0.08	0.06	0.21	0.00	0.00	0.01	0.00	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NPIQ elation or euphoria	0.09	0.08	0.06	0.21	0.00	0.00	0.01	0.00	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NPIQ apathy or indifference	0.10	0.08	0.06	0.21	0.00	0.00	0.01	0.00	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NPIQ disinhibition	0.09	0.08	0.06	0.21	0.00	0.00	0.01	0.00	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NPIQ irritability or lability	0.09	0.08	0.06	0.21	0.00	0.00	0.01	0.00	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NPIQ motor disturbance	0.10	0.08	0.06	0.21	0.00	0.00	0.01	0.00	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NPIQ nighttime behaviors	0.11	0.09	0.06	0.21	0.01	0.00	0.01	0.00	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NPIQ appetite	0.10	0.08	0.06	0.21	0.00	0.00	0.01	0.00	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FAQ bills	0.11	0.16	0.17	0.28	0.08	0.07	0.18	0.32	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FAQ taxes	0.13	0.20	0.22	0.37	0.10	0.15	0.17	0.27	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FAQ shopping	0.09	0.10	0.08	0.25	0.01	0.04	0.09	0.09	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FAQ games	0.13	0.17	0.23	0.33	0.10	0.15	0.19	0.36	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FAQ stove	0.09	0.09	0.08	0.26	0.02	0.00	0.02	0.05	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FAQ meal prep	0.11	0.17	0.19	0.40	0.07	0.15	0.22	0.23	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FAQ events	0.09	0.08	0.07	0.21	0.01	0.00	0.01	0.09	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FAQ pay attention	0.09	0.08	0.06	0.21	0.01	0.00	0.02	0.00	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FAQ remdates	0.09	0.08	0.06	0.20	0.00	0.00	0.01	0.05	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FAQ travel	0.09	0.08	0.07	0.21	0.01	0.00	0.01	0.05	0.03	0.04	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History family cognitive impairment	0.15	0.18	0.15	0.28	1.00	1.00	1.00	1.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History heart attack/cardiac arrest	0.35	0.16	0.12	0.24	0.00	0.00	0.01	0.00	0.11	0.07	0.05	0.89	0.81	0.94	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History atrial fibrillation	0.35	0.16	0.12	0.23	0.01	0.00	0.01	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History angioplasty/endarterectomy/stent	0.35	0.16	0.12	0.23	0.00	0.00	0.01	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History cardiac bypass procedure	0.35	0.16	0.12	0.23	0.00	0.00	0.01	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History pacemaker	0.35	0.16	0.12	0.24	0.00	0.00	0.01	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History congestive heart failure	0.35	0.16	0.12	0.25	0.00	0.00	0.01	0.00	1.00	1.00	1.00	0.92	0.95	1.00	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History other cardiovascular disease	0.35	0.16	0.12	0.23	0.00	0.00	0.02	0.00	1.00	1.00	1.00	0.96	0.99	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History stroke	0.35	0.16	0.12	0.24	0.00	0.00	0.02	0.05	0.00	0.00	0.00	0.92	0.92	0.82	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History transient ischemic attack	0.36	0.16	0.13	0.24	0.01	0.07	0.03	0.00	1.00	1.00	1.00	0.94	0.96	0.88	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History seizures	0.35	0.16	0.12	0.25	0.00	0.00	0.01	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History traumatic brain injury	0.36	0.16	0.14	0.23	0.02	0.00	0.01	0.05	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History hypertension	0.35	0.16	0.12	0.23	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.31	0.35	0.29	0.22	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History hypercholesterolemia	0.36	0.17	0.14	0.24	0.01	0.07	0.01	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History diabetes	0.36	0.16	0.12	0.23	0.00	0.00	0.00	0.00	1.00	1.00	1.00	0.86	0.84	0.82	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History vitamin B12 deficiency	0.37	0.18	0.14	0.24	0.04	0.00	0.04	0.05	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History Thyroid disease	0.36	0.17	0.13	0.23	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History incontinence-urinary	0.35	0.16	0.12	0.24	0.00	0.00	0.01	0.05	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History incontinence-bowel	0.35	0.16	0.12	0.23	0.00	0.00	0.01	0.05	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History active depression in the last two years	0.36	0.17	0.13	0.24	0.00	0.00	0.01	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History depression episodes more than two years ago	0.36	0.18	0.13	0.25	0.01	0.00	0.01	0.00	0.52	0.59	0.47	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History other psychiatric disorder	0.35	0.16	0.12	0.25	0.00	0.00	0.00	0.00	0.11	0.07	0.05	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History alcohol abuse	0.35	0.16	0.12	0.23	0.01	0.00	0.02	0.05	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History smoked more than 100 cigarettes in life	0.35	0.16	0.13	0.24	0.01	0.04	0.03	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History total years smoked cigarettes	0.37	0.18	0.15	0.27	0.57	0.67	0.57	0.73	0.84	0.80	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History average number of packs smoked per day	0.36	0.18	0.14	0.25	0.56	0.70	0.57	0.77	0.84	0.80	0.83	0.04	0.04	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
History other abused substances	0.35	0.16	0.12	0.24	0.00	0.00	0.01	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Figure B.1. Non-imaging features missing rate. The proportion of missing data is shown for all non-imaging features across the eight cohorts. A value of 0.0 represents that no data is missing, while a value of 1.0 indicates that all data for a particular feature was absent. We further stratify missingness by diagnostic label (NC, MCI, AD, and nADD) to demonstrate instances in which data-availability and disease status may be correlated.

Appendix C. Classification Performance between AD and stratified nADD

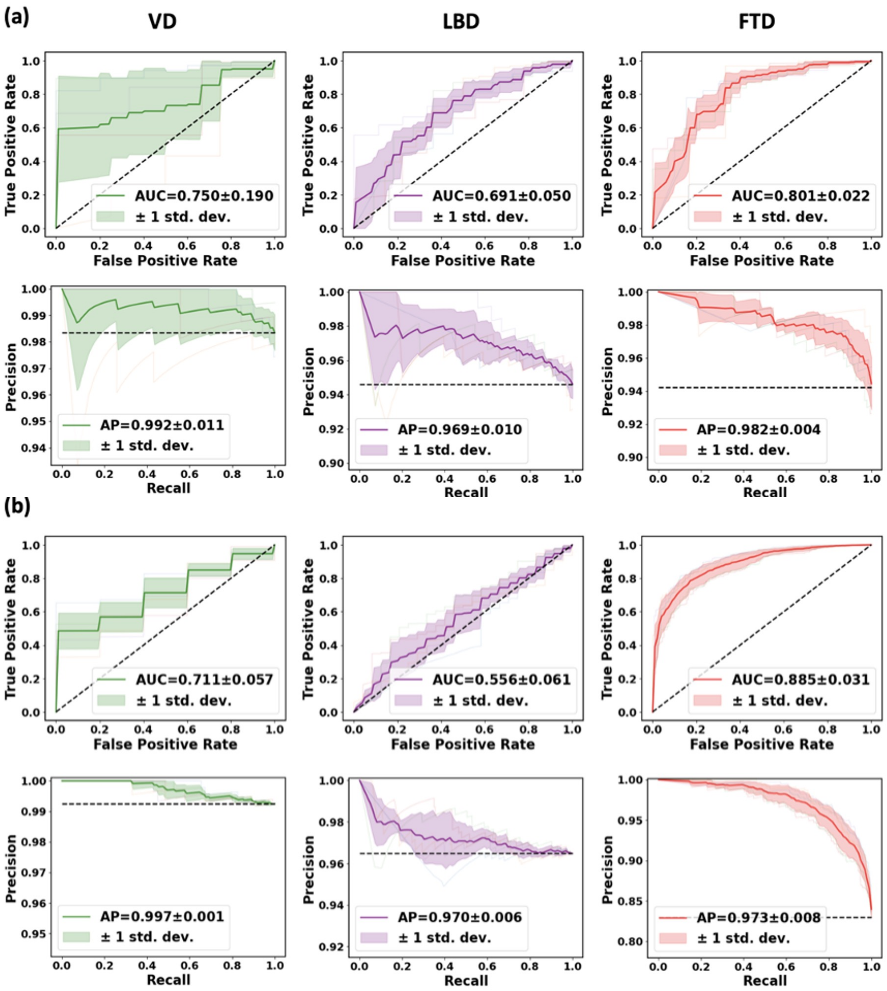


Figure C.1. MRI model performance on AD task across nADD sub-groups. The ROC and PR curves demonstrate the ability of the MRI-only model to accurately delineate AD from nADD cases when evaluated across several nADD subgroups, including vascular dementia (VD), Lewy body dementia (LBD), and frontotemporal dementia (FTD). The area under ROC and PR curves demonstrate that model performance is the strongest on non-Parkinsonian dementias (VD and FTD) than Parkinsonian dementias (PDD and LBD). Performance from the (a) NACC test set and (b) all external datasets are shown.

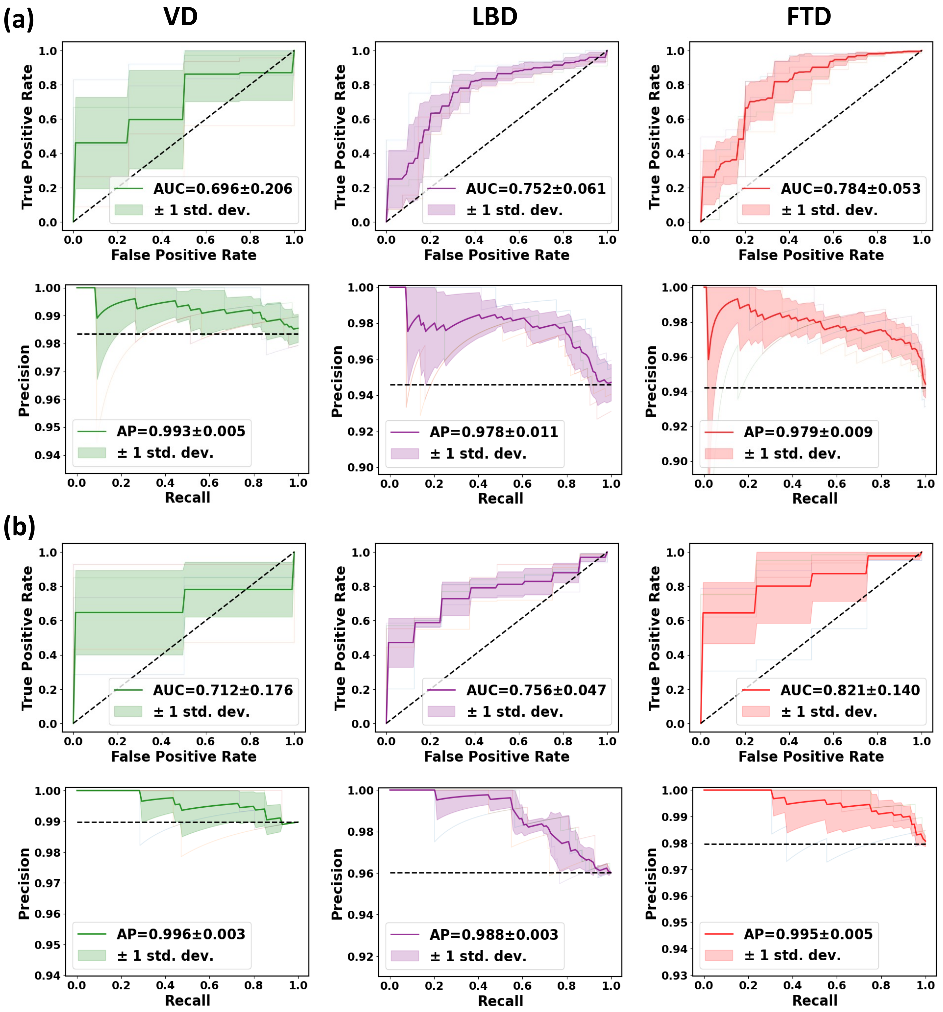


Figure C.2. Fusion model performance on AD task across nADD sub-groups. The ROC and PR curves demonstrate the ability of the fusion (CNN and CatBoost) model to accurately delineate AD from nADD cases when evaluated across the nADD subgroups. The area under ROC and PR curves demonstrate that the inclusion of non-imaging data elevates the model's performance on Parkinsonian dementias. Performance from the (a) NACC test set and (b) OASIS dataset is shown.

Appendix D. MRI Model's ROC & PR Curves on 7 Cohorts

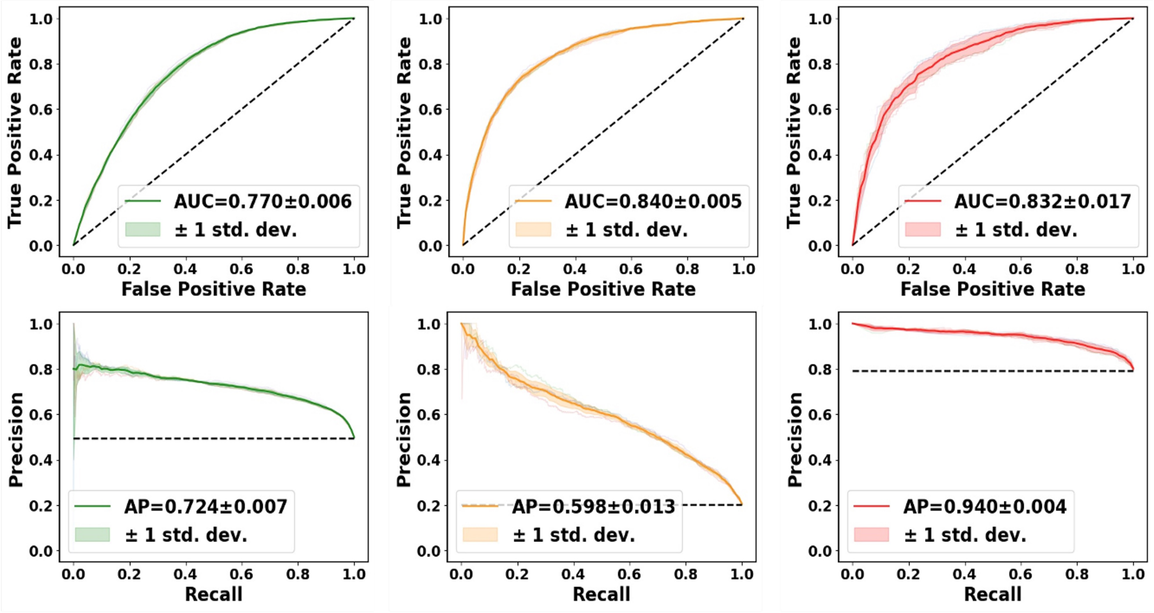


Figure D.1. MRI model's performance on all external data. ROC curves (row 1) on COG_{NC} task (left), COG_{DE} task (middle) and ADD task (right). PR curves (row 2) on COG_{NC} task (left), COG_{DE} task (middle) and ADD task (right).

Appendix E. Network Node Definition

index	Adult brain atlas [region]	Sagittal view [node index]	axial view [node index]
1	TL hippocampus R	21	37
2	TL hippocampus L		38
3	TL amygdala R	18	29
4	TL amygdala L		30
5	TL anterior temporal lobe medial part R	19	33
6	TL anterior temporal lobe medial part L		34
7	TL anterior temporal lobe lateral part R		31
8	TL anterior temporal lobe lateral part L		32
9	TL parahippocampal and ambient gyrus R	23	41
10	TL parahippocampal and ambient gyrus L		42
11	TL superior temporal gyrus middle part R		47
12	TL superior temporal gyrus middle part L		48
13	TL middle and inferior temporal gyrus R	22	39
14	TL middle and inferior temporal gyrus L		40
15	TL fusiform gyrus R	20	35
16	TL fusiform gyrus L		36
17	cerebellum R	27	
18	cerebellum L		
19	brainstem excluding substantia nigra	25	
20	insula posterior long gyrus L	29	
21	insula posterior long gyrus R		
22	OL lateral remainder occipital lobe L	12	26
23	OL lateral remainder occipital lobe R		25
24	CG anterior cingulate gyrus L	1	
25	CG anterior cingulate gyrus R		
26	CG posterior cingulate gyrus L	2	2
27	CG posterior cingulate gyrus R		1
28	FL middle frontal gyrus L	4	
29	FL middle frontal gyrus R		
30	TL posterior temporal lobe L		44
31	TL posterior temporal lobe R		43
32	PL angular gyrus L	14	
33	PL angular gyrus R		
34	caudate nucleus L	26	50
35	caudate nucleus R		49
36	nucleus accumbens L	30	
37	nucleus accumbens R		
38	putamen L	31	55
39	putamen R		54
40	Thalamus L	33	57
41	Thalamus R		56
42	Pallidum L		53
43	Pallidum R		52
44	Corpus callosum	28	51
45	Lateral ventricle excluding temporal horn R	9	

46	Lateral ventricle excluding temporal horn L		
47	Lateral ventricle temporal horn R	10	
48	Lateral ventricle temporal horn L		
49	Third ventricle	24	
50	FL precentral gyrus L	6	16
51	FL precentral gyrus R		15
52	FL straight gyrus L	7	18
53	FL straight gyrus R		17
54	FL anterior orbital gyrus L	3	4
55	FL anterior orbital gyrus R		3
56	FL inferior frontal gyrus L		6
57	FL inferior frontal gyrus R		5
58	FL superior frontal gyrus L	8	22
59	FL superior frontal gyrus R		21
60	PL postcentral gyrus L	15	
61	PL postcentral gyrus R		
62	PL superior parietal gyrus L	16	
63	PL superior parietal gyrus R		
64	OL lingual gyrus L	13	28
65	OL lingual gyrus R		27
66	OL cuneus L	11	24
67	OL cuneus R		23
68	FL medial orbital gyrus L	3	10
69	FL medial orbital gyrus R		9
70	FL lateral orbital gyrus L		8
71	FL lateral orbital gyrus R		7
72	FL posterior orbital gyrus L		12
73	FL posterior orbital gyrus R		11
74	substantia nigra L	32	
75	substantia nigra L		
76	FL subgenual frontal cortex L		20
77	FL subgenual frontal cortex R		19
78	FL subcallosal area L		
79	FL subcallosal area R		
80	FL pre-subgenual frontal cortex L	5	14
81	FL pre-subgenual frontal cortex R		13
82	TL superior temporal gyrus anterior part L		46
83	TL superior temporal gyrus anterior part R		45
84	PL supramarginal gyrus L	17	
85	PL supramarginal gyrus R		
86	insula anterior short gyrus L	29	
87	insula anterior short gyrus R		
88	Insula middle short gyrus L		
89	Insula middle short gyrus R		
90	insula posterior short gyrus L		
91	insula posterior short gyrus R		
92	insula anterior inferior cortex L		
93	insula anterior inferior cortex R		
94	insula anterior long gyrus L		
95	insula anterior long gyrus R		

Table E.1. Definition of network nodes. We presented the network visualization of the inter-region correlation structure of the brain from the axial and sagittal views. The node was defined to represent a particular region from the parcellated brain and the edges between nodes demonstrated the sign and degree of correlation between a pair of nodes. We parcellated the brain MRI into 95 regions using the segmentation mask provided from the Hammersmith Adult Brain Atlas. To project the 3D structure into an axial and sagittal plane, we redefined the node for best visualization purpose. The "index" and "Adult brain atlas" columns show the completed 95 structures and their corresponding indexes from the Hammersmith Adult Brain Atlas. The "sagittal view" and "axial view" columns demonstrate how we merged and re-indexed regions and the node index number is what we labeled each node. In the sagittal view, we focused on visualizing the correlation between the temporal lobe, frontal lobe, parietal lobe, occipital lobe, cerebellum, and brainstem. Specifically, we merged the same structures from the left and right hemisphere as a single node in the sagittal projection, thus ending up with a total of 33 final nodes as defined in this table. In the axial view, we excluded some of the structures that have been already shown in the sagittal view, for example, insula, the third ventricle, etc. The focus of the axial view is to reveal the correlation between cerebrum structures from the left and right hemispheres. Our selection of the axial nodes yielded 57 regions.

Appendix F. Model Performance with Different Feature Combinations

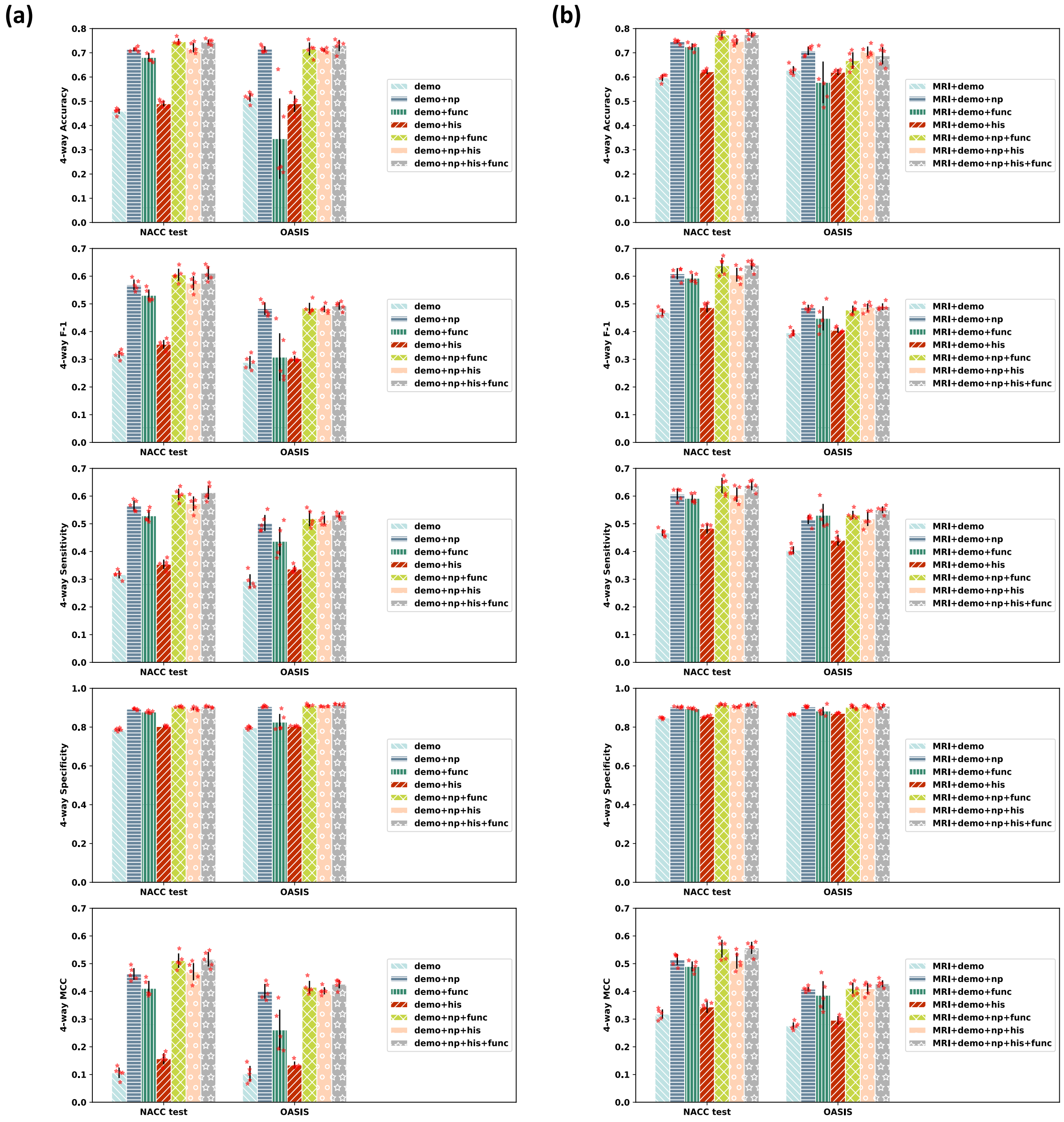


Figure F.1. Non-imaging and fusion models with partial features. We assessed the performance of 4-way classification on non-imaging and fusion (CNN + CatBoost) models when using varying combinations of historical (“his”), neuropsychological (“np”), and functional (“func”) variables (as well as MRI-derived variables in the case of the fusion model). The panel (a) on the left shows the models’ performance using different feature combinations but without MRI information. The panel (b) on the right shows the model performance using various feature combinations with MRI included. The model accuracy, F-1, sensitivity, specificity, and MCC values are demonstrated, and comparison is made between the NACC test set and the OASIS dataset for each performance

metric. Of note, similar distributions of performance metrics are observed between the two datasets, thus suggesting that the model does not privilege particular features in one dataset over the other.

Appendix G. Model vs Experts Performance

a

	COG	**COG_{NC}**	**COG_{MCI}**	**COG_{DE}**	**ADD**	**4-way**
Accuracy	0.713±0.093 [0.665-0.761]	0.849±0.067 [0.815-0.883]	0.744±0.067 [0.710-0.778]	0.834±0.070 [0.798-0.870]	0.624±0.080 [0.583-0.665]	0.565±0.070 [0.529-0.601]
F-1	0.678±0.089 [0.632-0.724]	0.739±0.064 [0.706-0.772]	0.486±0.123 [0.423-0.549]	0.811±0.107 [0.756-0.866]	0.605±0.100 [0.554-0.656]	0.549±0.074 [0.511-0.587]
Sensitivity	0.695±0.072 [0.658-0.732]	0.821±0.113 [0.763-0.879]	0.494±0.163 [0.410-0.578]	0.768±0.170 [0.681-0.855]	0.598±0.175 [0.508-0.688]	0.565±0.070 [0.529-0.601]
Specificity	0.861±0.039 [0.841-0.881]	0.858±0.113 [0.800-0.916]	0.827±0.080 [0.786-0.868]	0.899±0.062 [0.867-0.931]	0.649±0.192 [0.550-0.748]	0.855±0.023 [0.843-0.867]
MCC	0.556±0.113 [0.498-0.614]	0.657±0.083 [0.614-0.700]	0.325±0.162 [0.242-0.408]	0.685±0.119 [0.624-0.746]	0.262±0.163 [0.178-0.346]	0.429±0.091 [0.382-0.476]

b

	COG	**COG_{NC}**	**COG_{MCI}**	**COG_{DE}**	**ADD**	**4-way**
Accuracy	0.520±0.029 [0.484-0.556]	0.734±0.021 [0.708-0.760]	0.600±0.047 [0.542-0.658]	0.706±0.039 [0.658-0.754]	0.688±0.048 [0.628-0.748]	0.412±0.026 [0.380-0.444]
F-1	0.475±0.030 [0.438-0.512]	0.398±0.055 [0.330-0.466]	0.341±0.077 [0.245-0.437]	0.686±0.037 [0.640-0.732]	0.754±0.028 [0.719-0.789]	0.375±0.034 [0.333-0.417]
Sensitivity	0.479±0.029 [0.443-0.515]	0.360±0.091 [0.247-0.473]	0.432±0.146 [0.251-0.613]	0.644±0.067 [0.561-0.727]	0.952±0.030 [0.915-0.989]	0.412±0.026 [0.380-0.444]
Specificity	0.761±0.018 [0.739-0.783]	0.859±0.050 [0.797-0.921]	0.656±0.098 [0.534-0.778]	0.768±0.101 [0.643-0.893]	0.424±0.106 [0.292-0.556]	0.804±0.009 [0.793-0.815]
MCC	0.247±0.045 [0.191-0.303]	0.241±0.044 [0.186-0.296]	0.078±0.075 [-0.015-0.171]	0.421±0.083 [0.318-0.524]	0.444±0.082 [0.342-0.546]	0.236±0.044 [0.181-0.291]

c

	COG	**COG_{NC}**	**COG_{MCI}**	**COG_{DE}**	**ADD**	**4-way**
Accuracy	0.618±0.042 [0.566-0.670]	0.878±0.017 [0.857-0.899]	0.634±0.045 [0.578-0.690]	0.724±0.050 [0.662-0.786]	0.688±0.097 [0.568-0.808]	0.544±0.022 [0.517-0.571]
F-1	0.619±0.034 [0.577-0.661]	0.804±0.023 [0.775-0.833]	0.441±0.027 [0.407-0.475]	0.612±0.098 [0.490-0.734]	0.740±0.046 [0.683-0.797]	0.499±0.037 [0.453-0.545]
Sensitivity	0.675±0.024 [0.645-0.705]	1.000±0.000 [1.000-1.000]	0.576±0.054 [0.509-0.643]	0.448±0.100 [0.324-0.572]	0.872±0.111 [0.734-1.010]	0.544±0.022 [0.517-0.571]
Specificity	0.830±0.019 [0.806-0.854]	0.837±0.023 [0.808-0.866]	0.653±0.071 [0.565-0.741]	1.000±0.000 [1.000-1.000]	0.504±0.271 [0.168-0.840]	0.848±0.007 [0.839-0.857]

MCC	0.497±0.035 [0.454-0.540]	0.751±0.028 [0.716-0.786]	0.205±0.049 [0.144-0.266]	0.537±0.078 [0.440-0.634]	0.396±0.204 [0.143-0.649]	0.401±0.031 [0.363-0.439]

d

	COG	**COG_{NC}**	**COG_{MCI}**	**COG_{DE}**	**ADD**	**4-way**
Accuracy	0.632±0.071 [0.544-0.720]	0.874±0.039 [0.826-0.922]	0.650±0.068 [0.566-0.734]	0.740±0.053 [0.674-0.806]	0.680±0.100 [0.556-0.804]	0.558±0.061 [0.482-0.634]
F-1	0.631±0.070 [0.544-0.718]	0.802±0.051 [0.739-0.865]	0.448±0.092 [0.334-0.562]	0.642±0.095 [0.524-0.760]	0.746±0.053 [0.680-0.812]	0.505±0.065 [0.424-0.586]
Sensitivity	0.683±0.065 [0.602-0.764]	1.000±0.000 [1.000-1.000]	0.568±0.132 [0.404-0.732]	0.480±0.107 [0.347-0.613]	0.912±0.053 [0.846-0.978]	0.558±0.061 [0.482-0.634]
Specificity	0.836±0.031 [0.798-0.874]	0.832±0.052 [0.767-0.897]	0.677±0.071 [0.589-0.765]	1.000±0.000 [1.000-1.000]	0.448±0.238 [0.153-0.743]	0.853±0.020 [0.828-0.878]
MCC	0.510±0.087 [0.402-0.618]	0.748±0.064 [0.669-0.827]	0.221±0.145 [0.041-0.401]	0.562±0.083 [0.459-0.665]	0.407±0.167 [0.200-0.614]	0.409±0.075 [0.316-0.502]

Table G.1. Comparison of model performance with the neurologists. We randomly sampled 100 subjects from the NACC dataset. For each of the selected subject, we provided MRI scan along with a set of non-imaging features as specified in the supplementary material to 17 neurologists for them to review and make a prediction on one of the 4 possible categories, i.e., normal cognition (NC), mild cognitive impairment (MCI), Alzheimer's disease (AD), non-AD dementia (nADD). To make a head-to-head comparison, we also tested our MRI model, non-imaging model and fusion model on the same 100 selected subjects. We reported performance metrics, including accuracy, F-1, sensitivity, specificity, and Matthew correlation coefficient (MCC) for various tasks as indicated by column names based on the predictions from (a) 17 neurologists, (b) the MRI model, (c) the non-imaging model and (d) the fusion model. The mean and standard deviation (std) from table (a) was calculated over all 17 neurologists, and the mean and std from the other tables was derived from 5-fold validation experiments. More specifically, the COG represents the full classification of NC, MCI, and DE cases). In addition, we reported the performance of binary classification of NC vs. non-NC ("COG_{NC}" column), MCI vs. non-MCI ("COG_{MCI}" column) and DE vs. non-DE ("COG_{DE}" column). We also reported the model's performance in detecting AD from the dementied subjects within the "ADD columns. Lastly, we reported the 4-way classification of NC, MCI, AD, nADD ("4-way" column).

	Neuroradiologists	**MR-only model**
Accuracy	0.566±0.054 [0.516-0.616]	0.692±0.035 [0.649-0.735]
F-1	0.571±0.070 [0.506-0.636]	0.920±0.044 [0.865-0.975]
Sensitivity	0.589±0.122 [0.476-0.702]	0.464±0.090 [0.352-0.576]
Specificity	0.543±0.142 [0.412-0.674]	0.750±0.022 [0.723-0.777]
MCC	0.135±0.108 [0.035-0.235]	0.435±0.057 [0.364-0.506]

Table G.2. Comparison of model performance with the neuroradiologists. We randomly sampled 50 subjects from the NACC dataset. For each of the selected subject, we provided MRI scan along with a set of non-imaging features as specified in the supplementary material to 7 neuroradiologists for them to independently review and make a prediction on one of the 2 possible categories, i.e., Alzheimer's disease (AD) and non-AD dementia (nADD). We reported performance metrics, including accuracy, F-1, sensitivity, specificity, and Matthew correlation coefficient (MCC) for this binary classification task by considering AD as positive samples.

Appendix H. Classification Performance from Simple Thresholding

Variable	**COG_{NC} task AUC**	**COG_{NC} task AP**	**COG_{DE} task AUC**	**COG_{DE} task AP**	**ADD task AUC**	**ADD task AP**
trailA	0.783	0.79	0.817	0.587	0.52	0.877
trailB	0.818	0.839	0.853	0.564	0.532	0.869
boston	0.791	0.762	0.825	0.59	0.569	0.887
digitB	0.725	0.719	0.753	0.458	0.533	0.891
digitBL	0.704	0.69	0.735	0.413	0.522	0.884
digitF	0.66	0.649	0.684	0.383	0.528	0.881
digitFL	0.632	0.624	0.654	0.329	0.54	0.885
animal	0.839	0.824	0.878	0.702	0.501	0.869
gds	0.647	0.633	0.6	0.275	0.608	0.895
lm_imm	0.872	0.86	0.907	0.722	0.638	0.913
lm_del	0.895	0.886	0.916	0.706	0.713	0.93
mmse	0.881	0.848	0.931	0.814	0.616	0.896
npiq_DEL	0.545	0.543	0.58	0.339	0.522	0.871
npiq_HALL	0.526	0.533	0.544	0.294	0.55	0.878
npiq_AGIT	0.597	0.574	0.628	0.357	0.501	0.86
npiq_DEPD	0.588	0.57	0.6	0.301	0.523	0.872
npiq_ANX	0.608	0.582	0.642	0.348	0.539	0.877
npiq_ELAT	0.513	0.527	0.516	0.246	0.508	0.868
npiq_APA	0.623	0.59	0.67	0.417	0.58	0.887

npiq_DISN	0.556	0.55	0.569	0.299	0.566	0.882
npiq_IRR	0.603	0.578	0.607	0.321	0.52	0.87
npiq_MOT	0.559	0.551	0.589	0.338	0.528	0.873
npiq_NITE	0.567	0.554	0.577	0.307	0.552	0.878
npiq_APP	0.575	0.561	0.595	0.32	0.541	0.875
faq_BILLS	0.794	0.742	0.928	0.79	0.511	0.859
faq_TAXES	0.807	0.762	0.936	0.801	0.522	0.872
faq_SHOPPING	0.733	0.676	0.88	0.752	0.538	0.875
faq_GAMES	0.706	0.673	0.841	0.689	0.571	0.879
faq_STOVE	0.632	0.602	0.73	0.55	0.53	0.878
faq_MEALPREP	0.709	0.677	0.853	0.71	0.521	0.885
faq_EVENTS	0.75	0.687	0.867	0.723	0.54	0.874
faq_PAYATTN	0.736	0.674	0.846	0.684	0.518	0.872
faq_REMDATES	0.82	0.756	0.925	0.776	0.527	0.871
faq_TRAVEL	0.781	0.716	0.908	0.766	0.501	0.864

Table H.8. Classification performance of each standalone neuropsychological test. To compare our machine learning models to sample thresholding of common neuropsychiatric tests, we measured the area under the curve (AUC) of the ROC curves, and averaged precision (AP) of the precision-recall curve for the COG_{NC}, the COG_{DE} and the ADD tasks, respectively on the NACC cohort. The AUC and AP were derived by simply thresholding on each of the raw neuropsychiatric test score.

Bibliography

1. Rabins, P. V., Mace, N. L. & Lucas, M. J. The Impact of Dementia on the Family. *JAMA* **248**, 333–335 (1982).

2. Nichols, E. *et al.* Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **18**, 88–106 (2019).

3. Langa, K. M., Foster, N. L. & Larson, E. B. Mixed Dementia Emerging Concepts and Therapeutic Implications. *JAMA* **292**, 2901–2908 (2004).

4. James, B. D. *et al.* Contribution of Alzheimer disease to mortality in the United States. *Neurology* **82**, 1045–1050 (2014).

5. Association, A. 2019 Alzheimer's disease facts and figures. *Alzheimers Dement.* **15**, 321–387 (2019).

6. Selkoe, D. J. The molecular pathology of Alzheimer's disease. *Neuron* **6**, 487–498 (1991).

7. Lindsay, J. *et al.* Risk Factors for Alzheimer's Disease: A Prospective Analysis from the Canadian Study of Health and Aging. *Am. J. Epidemiol.* **156**, 445–453 (2002).

8. Tang, Y., Whitman, G. T., Lopez, I. & Baloh, R. W. Brain Volume Changes on Longitudinal Magnetic Resonance Imaging in Normal Older People. *J. Neuroimaging* **11**, 393–400 (2001).

9. Green, R. C. *et al.* Disclosure of APOE Genotype for Risk of Alzheimer's Disease. *N. Engl. J. Med.* **361**, 245–254 (2009).

10. Rajan, K. B. *et al.* Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020–2060). *Alzheimers Dement.* **17**, 1966–1975 (2021).

11. Hebert, L. E., Weuve, J., Scherr, P. A. & Evans, D. A. Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology* **80**, 1778–1783 (2013).

12. Warshaw, G. A. & Bragg, E. J. Preparing The Health Care Workforce To Care For Adults With Alzheimer's Disease And Related Dementias. *Health Aff. (Millwood)* **33**, 633–641 (2014).

13. Rafii, M. S. & Aisen, P. S. Recent developments in Alzheimer's disease therapeutics. *BMC Med.* **7**, 7 (2009).

14. Sevigny, J. *et al.* The antibody aducanumab reduces Aβ plaques in Alzheimer's disease. *Nature* **537**, 50–56 (2016).

15. Dubois, B., Padovani, A., Scheltens, P., Rossi, A. & Dell'Agnello, G. Timely Diagnosis for Alzheimer's Disease: A Literature Review on Benefits and Challenges. *J. Alzheimers Dis.* **49**, 617–631 (2016).

16. Mitchell, T. M. *Machine Learning*. (McGraw-Hill Education, 1997).

17. Hochreiter, S., Younger, A. S. & Conwell, P. R. Learning to Learn Using Gradient Descent. in *Artificial Neural Networks — ICANN 2001* (eds. Dorffner, G., Bischof, H. & Hornik, K.) 87–94 (Springer, 2001). doi:10.1007/3-540-44668-0_13.

18. De Jong, K. Learning with genetic algorithms: An overview. *Mach. Learn.* **3**, 121–138 (1988).

19. Rutenbar, R. A. Simulated annealing algorithms: an overview. *IEEE Circuits Devices Mag.* **5**, 19–26 (1989).

20. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).

21. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

22. Freund, Y. & Schapire, R. E. Large Margin Classification Using the Perceptron Algorithm. *Mach. Learn.* **37**, 277–296 (1999).

23. Gu, J. *et al.* Recent advances in convolutional neural networks. *Pattern Recognit.* **77**, 354–377 (2018).

24. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Phys. Nonlinear Phenom.* **404**, 132306 (2020).

25. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

26. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).

27. Kirkpatrick, J. *et al.* Pushing the frontiers of density functionals by solving the fractional electron problem. *Science* (2021) doi:10.1126/science.abj6511.

28. Stahl, B. C. Ethical Issues of AI. *Artif. Intell. Better Future* 35–53 (2021) doi:10.1007/978-3-030-69978-9_4.

29. Molnar, C., Casalicchio, G. & Bischl, B. Interpretable Machine Learning -- A Brief History, State-of-the-Art and Challenges. *ArXiv201009337 Cs Stat* **1323**, 417–431 (2020).

30. Definitions, methods, and applications in interpretable machine learning | PNAS. https://www.pnas.org/doi/10.1073/pnas.1900654116.

31. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).

32. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).

33. Ribeiro, M. T., Singh, S. & Guestrin, C. Anchors: High-precision model-agnostic explanations. in *Proceedings of the AAAI conference on artificial intelligence* vol. 32 (2018).

34. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).

35. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning Deep Features for Discriminative Localization. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2921–2929 (IEEE, 2016). doi:10.1109/CVPR.2016.319.

36. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. *ArXiv160204938 Cs Stat* (2016).

37. Qiu, S. *et al.* Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain* **143**, 1920–1933 (2020).

38. Philippe, B., Saad, Y. & Stewart, W. J. Numerical Methods in Markov Chain Modeling. *Oper. Res.* **40**, 1156–1179 (1992).

39. Taunk, K., De, S., Verma, S. & Swetapadma, A. A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* 1255–1260 (2019). doi:10.1109/ICCS45141.2019.9065747.

40. Noble, W. S. What is a support vector machine? *Nat. Biotechnol.* **24**, 1565–1567 (2006).

41. Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A. & Brown, S. D. An introduction to decision tree modeling. *J. Chemom.* **18**, 275–285 (2004).

42. Oshiro, T. M., Perez, P. S. & Baranauskas, J. A. How Many Trees in a Random Forest? in *Machine Learning and Data Mining in Pattern Recognition* (ed. Perner, P.) 154–168 (Springer, 2012). doi:10.1007/978-3-642-31537-4_13.

43. Aumüller, M., Bernhardsson, E. & Faithfull, A. ANN-Benchmarks: A Benchmarking Tool for Approximate Nearest Neighbor Algorithms. in *Similarity Search and Applications* (eds. Beecks, C., Borutta, F., Kröger, P. & Seidl, T.) 34–49 (Springer International Publishing, 2017). doi:10.1007/978-3-319-68474-1_3.

44. Dorogush, A. V., Ershov, V. & Gulin, A. CatBoost: gradient boosting with categorical features support. *ArXiv181011363 Cs Stat* (2018).

45. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016). doi:10.1145/2939672.2939785.

46. Wang, S.-C. Artificial Neural Network. in *Interdisciplinary Computing in Java Programming* (ed. Wang, S.-C.) 81–100 (Springer US, 2003). doi:10.1007/978-1-4615-0377-4_5.

47. Rosenblatt, F. The perceptron, a perceiving and recognizing automaton. *Cornell Aeronautical Laboratory* (1957).

48. Agarap, A. F. Deep Learning using Rectified Linear Units (ReLU). *ArXiv180308375 Cs Stat* (2019).

49. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4**, 251–257 (1991).

50. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *undefined* (1989).

51. Girosi, F., Jones, M. & Poggio, T. Regularization Theory and Neural Networks Architectures. *Neural Comput.* **7**, 219–269 (1995).

52. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv14091556 Cs* (2015).

53. Szegedy, C. *et al.* Going Deeper with Convolutions. *ArXiv14094842 Cs* (2014).

54. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems* vol. 25 (Curran Associates, Inc., 2012).

55. Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks. *ArXiv13112901 Cs* (2013).

56. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *ArXiv151203385 Cs* (2015).

57. Shelhamer, E., Long, J. & Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 640–651 (2017).

58. Huber, P. J. Robust Estimation of a Location Parameter. *Ann. Math. Stat.* **35**, 73–101 (1964).

59. Mannor, S., Peleg, D. & Rubinstein, R. The cross entropy method for classification. in *Proceedings of the 22nd international conference on Machine learning* 561–568 (Association for Computing Machinery, 2005). doi:10.1145/1102351.1102422.

60. Wang, Q., Ma, Y., Zhao, K. & Tian, Y. A Comprehensive Survey of Loss Functions in Machine Learning. *Ann. Data Sci.* **9**, 187–212 (2022).

61. Mitchell, M. *An Introduction to Genetic Algorithms*. (A Bradford Book, 1996).

62. Granville, V., Krivanek, M. & Rasson, J.-P. Simulated annealing: a proof of convergence. *IEEE Trans. Pattern Anal. Mach. Intell.* **16**, 652–656 (1994).

63. Dauphin, Y. *et al.* Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *ArXiv14062572 Cs Math Stat* (2014).

64. Ruder, S. An overview of gradient descent optimization algorithms. *ArXiv160904747 Cs* (2017).

65. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs* (2017).

66. Zeiler, M. D. ADADELTA: An Adaptive Learning Rate Method. *ArXiv12125701 Cs* (2012).

67. von Luxburg, U. & Schoelkopf, B. Statistical Learning Theory: Models, Concepts, and Results. *ArXiv08104752 Math Stat* (2008).

68. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

69. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv150203167 Cs* (2015).

70. Santurkar, S., Tsipras, D., Ilyas, A. & Madry, A. How Does Batch Normalization Help Optimization? in *Advances in Neural Information Processing Systems* vol. 31 (Curran Associates, Inc., 2018).

71. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **3**, e745–e750 (2021).

72. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

73. Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *ArXiv13096392 Stat* (2014).

74. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. *ArXiv170402685 Cs* (2019).

75. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. *ArXiv170301365 Cs* (2017).

76. Verma, S., Dickerson, J. & Hines, K. Counterfactual Explanations for Machine Learning: A Review. *ArXiv201010596 Cs Stat* (2020).

77. SONG, Y. & LU, Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch. Psychiatry* **27**, 130–135 (2015).

78. Yuan, Y., Wu, L. & Zhang, X. Gini-Impurity Index Analysis. *IEEE Trans. Inf. Forensics Secur.* **16**, 3154–3169 (2021).

79. Shapley, L. S. *Notes on the N-Person Game — II: The Value of an N-Person Game*. https://www.rand.org/pubs/research_memoranda/RM0670.html (1951).

80. Fisher, A., Rudin, C. & Dominici, F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *ArXiv180101489 Stat* (2019).

81. Gandomkar, Z., Khong, P. L., Punch, A. & Lewis, S. Using Occlusion-Based Saliency Maps to Explain an Artificial Intelligence Tool in Lung Cancer Screening: Agreement Between Radiologists, Labels, and Visual Prompts. *J. Digit. Imaging* (2022) doi:10.1007/s10278-022-00631-w.

82. Yee, E., Popuri, K., Beg, M. F. & Initiative, the A. D. N. Quantifying brain metabolism from FDG-PET images into a probability of Alzheimer's dementia score. *Hum. Brain Mapp.* **41**, 5–16 (2020).

83. Feng, X., Provenzano, F. A., Small, S. A., & for the Alzheimer’s Disease Neuroimaging Initiative. A deep learning MRI approach outperforms other biomarkers of prodromal Alzheimer’s disease. *Alzheimers Res. Ther.* **14**, 45 (2022).

84. Irie, R. *et al.* A Novel Deep Learning Approach with a 3D Convolutional Ladder Network for Differential Diagnosis of Idiopathic Normal Pressure Hydrocephalus and Alzheimer’s Disease. *Magn. Reson. Med. Sci.* **advpub**, mp.2019-0106 (2020).

85. Chattopadhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* 839–847 (2018). doi:10.1109/WACV.2018.00097.

86. Omeiza, D., Speakman, S., Cintas, C. & Weldermariam, K. Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models. *ArXiv190801224 Cs* (2019).

87. Wang, H. *et al.* Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. in 24–25 (2020).

88. Desai, S. & Ramaswamy, H. G. Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization. in 983–991 (2020).

89. Fu, R. *et al.* Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs. (2020) doi:10.48550/arXiv.2008.02312.

90. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for Simplicity: The All Convolutional Net. (2014) doi:10.48550/arXiv.1412.6806.

91. Jack, C. R. *et al.* Tracking pathophysiological processes in Alzheimer’s disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol.* **12**, 207–216 (2013).

92. Nordberg, A. PET imaging of amyloid in Alzheimer’s disease. *Lancet Neurol.* **3**, 519–527 (2004).

93. Mattsson, N. *et al.* Predicting diagnosis and cognition with 18F-AV-1451 tau PET and structural MRI in Alzheimer’s disease. *Alzheimers Dement. J. Alzheimers Assoc.* **15**, 570–580 (2019).

94. McKhann, G. M. *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. *Alzheimers Dement. J. Alzheimers Assoc.* **7**, 263–269 (2011).

95. Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P. & Thompson, P. M. The clinical use of structural MRI in Alzheimer disease. *Nat. Rev. Neurol.* **6**, 67–77 (2010).

96. Beach, T. G., Monsell, S. E., Phillips, L. E. & Kukull, W. Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005-2010. *J. Neuropathol. Exp. Neurol.* **71**, 266–273 (2012).

97. Qiu, S. *et al.* Fusion of deep learning models of MRI scans, Mini-Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment. *Alzheimers Dement. Amst. Neth.* **10**, 737–749 (2018).

98. Castelvecchi, D. Can we open the black box of AI? *Nature* **538**, 20–23 (2016).

99. Petersen, R. C. *et al.* Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology* **74**, 201–209 (2010).

100. Ellis, K. A. *et al.* Addressing population aging and Alzheimer's disease through the Australian imaging biomarkers and lifestyle study: collaboration with the Alzheimer's Disease Neuroimaging Initiative. *Alzheimers Dement. J. Alzheimers Assoc.* **6**, 291–296 (2010).

101. Massaro, J. M. *et al.* Managing and analysing data from a large-scale study on Framingham Offspring relating brain structure to cognitive function. *Stat. Med.* **23**, 351–367 (2004).

102. Beekly, D. L. *et al.* The National Alzheimer's Coordinating Center (NACC) Database: an Alzheimer disease database. *Alzheimer Dis. Assoc. Disord.* **18**, 270–277 (2004).

103. van de Pol, L. A. *et al.* Hippocampal atrophy in Alzheimer disease: age matters. *Neurology* **66**, 236–238 (2006).

104. Raji, C. A., Lopez, O. L., Kuller, L. H., Carmichael, O. T. & Becker, J. T. Age, Alzheimer disease, and brain structure. *Neurology* **73**, 1899–1905 (2009).

105. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

106. Dall, T. M. *et al.* Supply and demand analysis of the current and future US neurology workforce. *Neurology* **81**, 470–478 (2013).

107. Dall, T. M. Physician Workforce Shortages: What Do the Data Really Say? *Acad. Med.* **90**, 1581–1582 (2015).

108. Lu, D., Popuri, K., Ding, G. W., Balachandar, R. & Beg, M. F. Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET images. *Sci. Rep.* **8**, 5697 (2018).

109. Wada, A. *et al.* Differentiating Alzheimer's Disease from Dementia with Lewy Bodies Using a Deep Learning Technique Based on Structural Brain Connectivity. *Magn. Reson. Med. Sci.* **18**, 219–224 (2018).

110. Ma, D. *et al.* Differential Diagnosis of Frontotemporal Dementia, Alzheimer's Disease, and Normal Aging Using a Multi-Scale Multi-Type Feature Generative Adversarial Deep Neural Network on Structural Magnetic Resonance Images. *Front. Neurosci.* **14**, (2020).

111. Boxer, A. L. *et al.* Frontotemporal degeneration, the next therapeutic frontier: molecules and animal models for frontotemporal degeneration drug development. *Alzheimers Dement. J. Alzheimers Assoc.* **9**, 176–188 (2013).

112. Marek, K. *et al.* The Parkinson Progression Marker Initiative (PPMI). *Prog. Neurobiol.* **95**, 629–635 (2011).

113. Ellis, K. A. *et al.* The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int. Psychogeriatr.* **21**, 672–687 (2009).

114. LaMontagne, P. J. *et al.* OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. 2019.12.13.19014902 (2019) doi:10.1101/2019.12.13.19014902.

115. Mahmood, S. S., Levy, D., Vasan, R. S. & Wang, T. J. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet Lond. Engl.* **383**, 999–1008 (2014).

116. Linortner, P. *et al.* White Matter Hyperintensities Related to Parkinson's Disease Executive Function. *Mov. Disord. Clin. Pract.* **7**, 629–638 (2020).

117. Wang, F., Kaushal, R. & Khullar, D. Should Health Care Demand Interpretable Artificial Intelligence or Accept 'Black Box' Medicine? *Ann. Intern. Med.* **172**, 59–60 (2020).

118. Albert, M. *et al.* The Use of MRI and PET for Clinical Diagnosis of Dementia and Investigation of Cognitive Impairment: A Consensus Report. 15.

119. Kara, S. *et al.* Guidelines, training and quality assurance: influence on general practitioner MRI referral quality. *J. Prim. Health Care* **11**, 235–242 (2019).

120. Bernstein, A. *et al.* Dementia assessment and management in primary care settings: a survey of current provider practices in the United States. *BMC Health Serv. Res.* **19**, 919 (2019).

121. Zekry, D., Hauw, J.-J. & Gold, G. Mixed dementia: epidemiology, diagnosis, and treatment. *J. Am. Geriatr. Soc.* **50**, 1431–1438 (2002).

122. Graff-Radford, J. *et al.* New insights into atypical Alzheimer's disease in the era of biomarkers. *Lancet Neurol.* **20**, 222–234 (2021).

123. Wind, A. W. *et al.* Limitations of the Mini-Mental State Examination in diagnosing dementia in general practice. *Int. J. Geriatr. Psychiatry* **12**, 101–108 (1997).

124. McKeith, I. G. *et al.* Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium. *Neurology* **89**, 88–100 (2017).

125. Winblad, B. *et al.* Mild cognitive impairment--beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. *J. Intern. Med.* **256**, 240–246 (2004).

126. Chêne, G. *et al.* Gender and incidence of dementia in the Framingham Heart Study from mid-adult life. *Alzheimers Dement. J. Alzheimers Assoc.* **11**, 310–320 (2015).

127. Rascovsky, K. *et al.* Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain* **134**, 2456–2477 (2011).

128. Gorno-Tempini, M. L. *et al.* Classification of primary progressive aphasia and its variants. *Neurology* **76**, 1006–1014 (2011).

Curriculum vitae

Shangran Qiu

Education

2022	Ph.D. in Physics Boston University, Boston, MA
2016	B.S. in Physics Xi’an Jiaotong University, Xi’an, Shaanxi province, China

Areas of Interest

Computer vision methods for medical image analysis

Machine learning frameworks for disease diagnosis

Medicine-specific interpretability methods for machine learning models

Validation of machine learning frameworks with medical domain knowledge

Teaching Experience

2017	Boston University, Department of Physics Teaching Assistant PY211: some of the basic principles of physics, including forces, motion, momentum, energy, harmonic motion etc.
2016	Boston University, Department of Physics Teaching Assistant PY105: calculus-based introduction to basic principles physics with focus on Newtonian mechanics and conservation law.

Industrial Experience

2021	Facebook, Inc PhD machine learning internship Worked on improving the recommendation system for video contents in Facebook platform.
2019	Philips Research, North America Research Internship Worked on AI-based image analysis systems for echocardiogram data

Honors & Awards

2021	Alvaro Roccaro Memorial Prize
2021	Toffler Scholars in Neuroscience Award
2020	Boston University's Top 5 Alzheimer's Research Breakthroughs
2019	Boston University Data Science Day Brilliant Award
2014	Xi'an Jiaotong University Elite Student Award
2014	Siyuan Scholarship
2013	Pengkang Scholarship

Publications

[†] = equal contribution

1 Shangran Qiu[†], Matthew I. Miller[†], Prajakta S. Joshi, Joyce C. Lee, Chonghua Xue, Yunruo Ni, Yuwei Wang et al. "*Multimodal deep learning for Alzheimer's disease dementia assessment*." Nature Communications 13, 3404 (2022)

2 Romano, Michael F., Akshara Balachandra, Xiao Zhou, Michalina Jadick, Shangran Qiu, Diya Nijhawan, Sang P. Chin, Rhoda Au, and Vijaya B. Kolachalama. "*Comparative analysis of cerebrospinal fluid markers and multimodal imaging in predicting Alzheimer's disease progression.*" Alzheimer's & Dementia 17, 054457 (2021).

3 Zhou, Xiao[†], Shangran Qiu[†], Prajakta S. Joshi, Chonghua Xue, Ronald J. Killiany, Asim Z. Mian, Sang P. Chin, Rhoda Au, and Vijaya B. Kolachalama. "*Enhancing magnetic resonance imaging-driven Alzheimer's disease classification performance using generative adversarial learning*." Alzheimer's research & therapy 13, no. 1 (2021).

4 Qiu, Shangran[†], Prajakta S. Joshi[†], Matthew I. Miller[†], Chonghua Xue[†], Xiao Zhou, Cody Karjadi, Gary H. Chang et al. "*Development and validation of an interpretable deep*

learning framework for Alzheimer's disease classification." Brain 143, no. 6 (2020).

6 Chang, Gary H., David T. Felson, Shangran Qiu, Ali Guermazi, Terence D. Capellini, and Vijaya B. Kolachalama. "*Assessment of knee pain from MR imaging using a convolutional Siamese network*." European radiology 30, no. 6 (2020).

7 Qiu, Shangran, Megan S. Heydari, Matthew I. Miller, Prajakta S. Joshi, Benjamin C. Wong, Rhoda Au, and Vijaya B. Kolachalama. "*P1-119: Enhancing deep learning model performance for AD diagnosis using ROI-based selection*." Alzheimer's & Dementia 15 (2019).

8 Wang, Xiao, Quan Zhou, Jacob Harer, Gavin Brown, Shangran Qiu, Zhi Dou, John Wang, Alan Hinton, Carlos Aguayo Gonzalez, and Peter Chin. "*Deep learning-based classification and anomaly detection of side-channel signals*." In Cyber Sensing 2018, 10630, (2018).

9 Qiu, Shangran, Gary H. Chang, Marcello Panagia, Deepa M. Gopal, Rhoda Au, and Vijaya B. Kolachalama. "*Fusion of deep learning models of MRI scans, Mini–Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment*." Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 10 (2018).

10 Zhang, Yingchao, Dmitri V. Voronine, Shangran Qiu, Alexander M. Sinyukov, Mary Hamilton, Zachary Liege, Alexei V. Sokolov, Zhenrong Zhang, and Marlan O. Scully. "*Improving resolution in quantum subnanometre-gap tip-enhanced Raman nanoimaging*." Scientific reports 6, no. 1 (2016).