PNAS
www.pnas.org

## Main Manuscript for

# A Brain Imaging-Based Deep Learning Classifier for Sex and Alzheimer's Disease

Bin Lu[1,2], Hui-Xian Li[1,2], Zhi-Kai Chang[1,2], Le Li[3], Ning-Xuan Chen[1,2], Zhi-Chen Zhu[1,2], Hui-Xia Zhou[1,2], Zhen Fan[4], Hong Yang[5], Xiao Chen[1,2], Chao-Gan Yan[1,2,6,7]*, for the Alzheimer's Disease Neuroimaging Initiative**

[1]CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing, China; [2]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China; [3]Center for Cognitive Science of Language, Beijing Language and Culture University, Beijing, China; [4]Department of Neurosurgery, Huashan Hospital, Fudan University, Shanghai Neurosurgical Clinical Center, Shanghai, China; [5]Department of Radiology, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, Zhejiang, China; [6]International Big-Data Research Center for Depression (IBRCD), Institute of Psychology, Chinese Academy of Sciences, Beijing, China; [7]Magnetic Resonance Imaging Research Center, Institute of Psychology, Chinese Academy of Sciences, Beijing, China. *e-mail: ycg.yan@gmail.com. **Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

* Chao-Gan Yan

**Email:** ycg.yan@gmail.com

**Classification**

Biological Sciences, Neuroscience.

Social Sciences, Psychological and Cognitive Sciences

**Keywords**

Alzheimer's disease, convolutional neural network, magnetic resonance brain imaging, sex difference, transfer learning

**Author Contributions**

C.-G.Y. designed the overall experiment. B.L., H.-X.L., L.L., C.-N.X., Z.-H.C., H.-X.L., Z.F., H.Y. and X.C. applied and preprocessed imaging data. H.-X.L. and B.L sorted the phenotype information of datasets. B.L. designed the model architectures and trained the models, Z.-K.C, B.L. and C.-G.Y. built the online classifiers. C.G.Y. provided technical supports and supervised the project. B.L. and C.G.Y. wrote the paper.

**This PDF file includes:**

Main Text
Figures 1 to 5

**Abstract**

Beyond detecting brain damage or tumors with magnetic resonance brain imaging, little success has been attained in identifying individual differences, e.g., sex or brain disorders. The current study aims to build an industrial-grade brain imaging-based classifier to infer individual differences using deep learning/transfer learning on unprecedented big data. We pooled 34 datasets to constitute the largest brain magnetic resonance image sample to date (85,721 samples from 50,876 participants), and then applied a state-of-the-art deep convolutional neural network, Inception-ResNet-V2, to build an industrial-grade sex classifier. We achieved 94.9% accuracy in cross-dataset-validation, i.e., the model can classify the sex of a participant with brain structural imaging data from anybody and any scanner with about 95% accuracy. We then explored the potential of a deep convolutional network to objectively diagnose brain disorders. Using transfer learning, the model fine-tuned to Alzheimer's Disease (AD) achieved 88.4% accuracy in leave-sites-out five-fold cross-validation on the ADNI dataset and 86.1% accuracy for a direct test on an unseen independent dataset (OASIS). When directly testing this AD classifier on brain images of unseen mild cognitive impairment (MCI) patients, 63.2% who finally converted into AD were predicted as AD, versus 22.1% who did not convert into AD were predicted as AD. Predicted scores of the AD classifier showed significant correlations with severity of illness. The transfer learning framework failed to achieve practical accuracy for psychiatric disorders. Our trained model, code and preprocessed data were openly-shared to demonstrate how open-science and deep learning can advance the clinical utility of brain imaging.

**Significance Statement**

The accumulation of open-access brain-imaging datasets makes it possible for applying deep learning on magnetic resonance brain-imaging in identifying individual differences. We pooled 34 brain-imaging datasets to constitute the largest sample to date and built a model that can classify the sex of a participant from brain structural imaging from anybody and any scanner with about 95% accuracy. Using transfer learning, the model fine-tuned to Alzheimer's Disease (AD) achieved 88% accuracy on ADNI dataset and achieved 86% accuracy on an unseen independent dataset. It also showed practical utility in predicting the progression of mild cognitive impairment patients. Our model, code and preprocessed data were openly-shared to facilitate open-science and deep learning in clinical utility of brain imaging.

## Main Text

### Introduction

Can we infer individual differences from brain information? This is a question that has been long pursued. Brain imaging provides a tool to achieve this goal non-invasively. However, beyond the achievement of visually identifying brain damage or tumors, little success has been attained in identifying individual differences, e.g., age, sex, personality, or mental illness. This is challenging since humans cannot discern the relevant features from brain images, while machine intelligence may provide a way. The current study aims to build an industrial-grade brain imaging-based classifier to infer individual differences by utilizing deep learning/transfer learning on big data.

Progress has been made in using brain imaging, especially magnetic resonance imaging (MRI), to predict sex (1, 2), age (3, 4), personality (5), Alzheimer's Disease (AD) (6, 7), major depressive disorder (MDD) (8, 9), attention-deficit/hyperactivity disorder (ADHD) (10), and autism spectrum disorder (ASD) (11, 12) among others. However, all of these efforts have failed to generalize. Brain imaging data varies depending on scanner characteristics such as scanner vendor, head coil, sequence, applied gradient fields, reconstruction methods, scanning parameters, voxel size, field of view, etc. Participant characteristics also vary in sex, age, race and education, etc. These variations make a brain imaging-based classifier trained on a site (or several sites) difficult to generalize to unseen sites/scanners, thus preventing brain imaging-based classifiers from becoming practically useful, e.g., in clinical settings.

Recently, utilizing deep learning on big data has been successfully applied on an industrial-grade in fields like extreme weather condition prediction (13), aftershock pattern prediction (14) and automatic speech recognition (15). In medical imaging, image-based deep convolutional neural networks (CNN) have been applied to objectively diagnose retinal diseases (16), skin cancer (17) and breast cancer screening (18). In brain imaging, CNN have predicted chronological age with high accuracy (3, 4). However, accuracy has been insufficient when generalizing to unseen datasets (i.e., for data acquired in difference sites/scanners, Pearson's correlation coefficients between predicted and actual age range from 0.53 to 0.86 (3)). Brain age prediction errors may be biologically meaningful as brain disorders may involve accelerated or delayed brain maturation/aging (3). Nonetheless, a brain imaging-based CNN classifier has yet to achieve practical utility.

A bottleneck for developing an industrial-grade brain imaging-based classifier is the scale and the variety of the training datasets. In recent years, data sharing projects have made upwards of 100,000 brain images available to the scientific community. However, no studies have fully implemented this resource to train classifiers. The largest training sample (n = 45,615) came mainly from a single site (UKBiobank) (3). The second and third largest training data sets comprised 16,848 (4) and 11,271 (19) participants. Even with a relatively large sample, if the training sample doesn't contain many sites (i.e., has variations in manufacturers of MR equipment, scanning parameters, quality control procedures and characteristics of participants, etc.), a trained classifier will fail to generalize to unseen datasets. Thus, in the current study, we utilized the largest sample to date (n = 85,721 from 34 datasets, see Table S1), to achieve an industrial-grade classifier which can generalize to any scanner and any sample from anywhere in the world.

Our primary training goal was sex prediction, as this is an objective dichotomous indicator available for every participant in open datasets. After obtaining an industrial-grade brain imaging-based classifier for sex with high cross-dataset accuracy, we used transfer learning to see if we could classify patients with AD, MDD, ASD or ADHD. As these disorder datasets only contain several thousand patients and corresponding healthy controls, direct training on these datasets may result in overfitting with poor generalization to new unseen testing data. A transfer learning framework was proposed to solve this problem, by training a model on a certain characteristic based on big

data samples and fine-tuning it to another characteristic in smaller samples (20), following successful examples in diagnosing retinal disease (16) and skin cancer (17). To attain generalizability, we used cross-site validation to make sure the accuracy of the classifier would be insensitive to site/scanner variability. We openly share our preprocessed data, trained model, code and framework, and have built an online predicting website (http://brainimagenet.org:8088) for anyone interested in testing our classifier with brain imaging data from anybody and any scanner.

## Results

### Brain imaging big data

Only brain imaging data with enough size and variety can make deep learning useful for building an industrial-grade classifier. We submitted data access applications to nearly all the open-access brain imaging data archives, and received permissions from the administrators of 34 datasets (85,721 samples from 50,876 participants, see Table S1, no application requirements for some datasets). Data for each participant contained at least one session with a T1-weighted brain structural image and information on participant sex.

### Performance of the sex classifier

We trained a 3-dimension Inception-ResNet-v2 model adopted from its 2-dimension version in the Keras built-in application (see Fig. 1*A* for structure). This is a record-breaking model in pattern recognition (21) which integrates two series of classical CNN models, Inception and ResNet. We replaced the convolution, pooling and normalization modules with their 3-dimension versions and adjusted the number of layers and convolutional kernels to make them suitable for 3-dimension MRI inputs. We did not feed raw data to the classifier for training, but used prior knowledge in brain imaging data analysis. The brain structural data were segmented and normalized to yield grey matter density (GMD) and grey matter volume (GMV) maps (e.g., GMD and GMV maps were treated as different input channels). The present model can take advantage of convolutional kernels with different sizes and shapes and extract features in different sizes, and mitigate vanishing gradients and exploding gradients by adding residual modules.

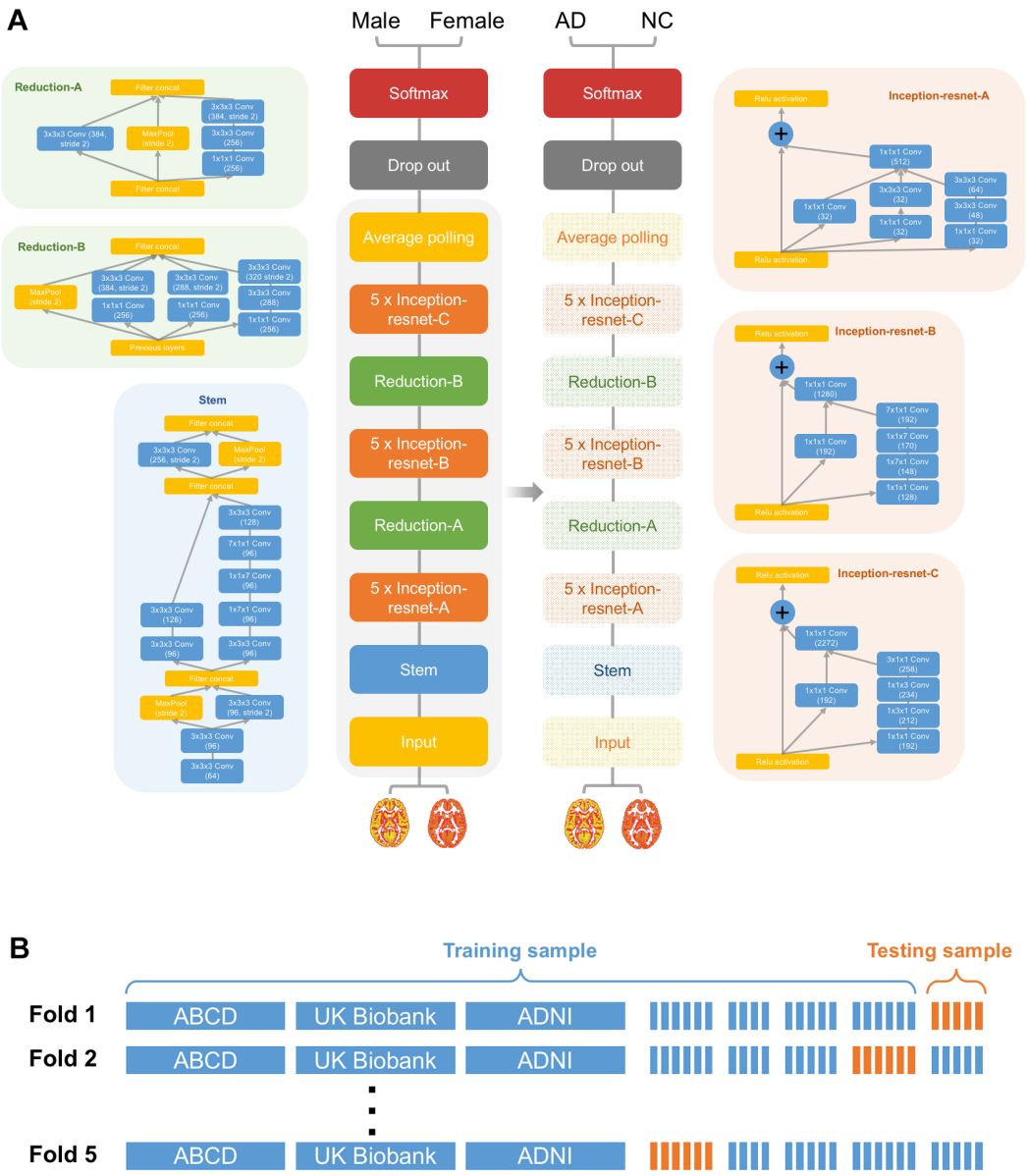


**Fig. 1.** Flow diagram for the training procedure for the sex classifier and transfer learning framework. *(A)* Schema for 3D Inception-ResNet-V2 network and transfer learning framework. *(B)* Schematic diagram for leave-dataset-out 5-fold cross-validation in training the sex classifier.

To ensure generalizability, five-fold cross-dataset validation was used to assess classifier accuracy (see Fig. 1*B*). Of note, 3 datasets were always kept in the training sample due to their massive number of samples: Adolescent Brain Cognitive Development Study (ABCD, n = 31,176), UK Biobank (n = 20,124) and Alzheimer's Disease Neuroimaging Initiative (ADNI, n = 16,596). The five-fold cross-dataset validation accuracies are: 94.8%, 94.0%, 94.8%, 95.7% and 95.8%. Taking together, accuracy is 94.9% in testing samples when pooling results in five folds. The area under curve (AUC) of the receiver operating characteristic (ROC) curve reached 0.981 (see Fig. 2). In short, our model can classify the sex of a participant with brain structural imaging data from anyone and any scanner with an accuracy of about 95%. Interested readers can test this model on our

online predicting website (http://brainimagenet.org:8088). The code and model are also openly shared at https://github.com/Chaogan-Yan/BrainImageNet.

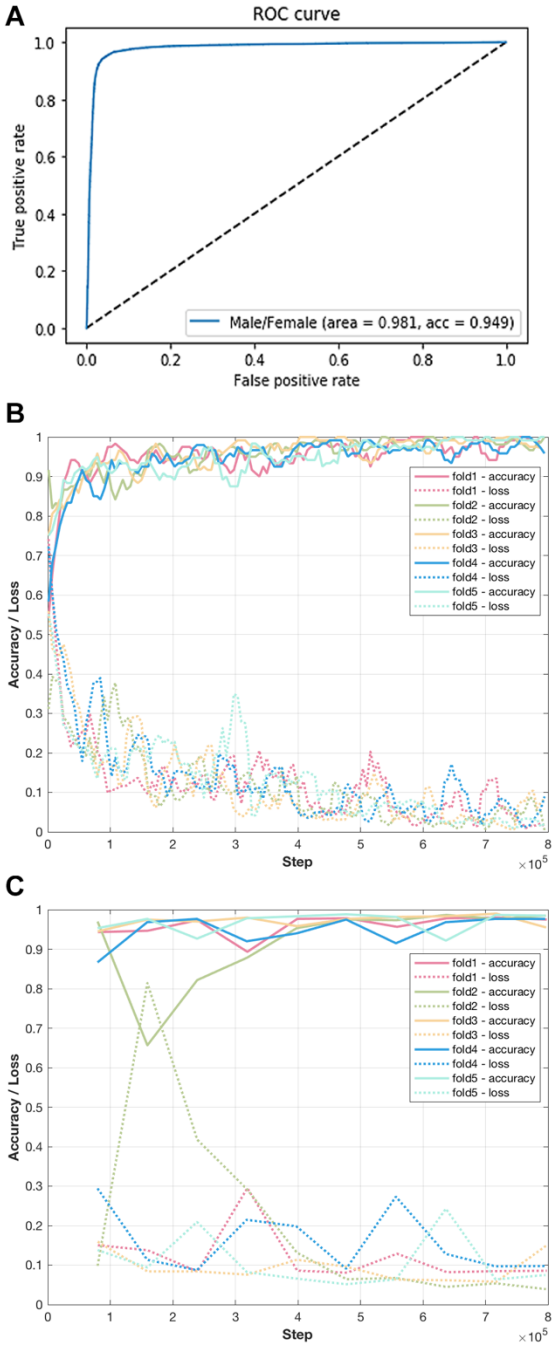


**Fig. 2.** Performance of the sex classifier. *(A)* The receiver operating characteristic curve of the sex classifier. *(B)* The tensorboard monitor graph of the sex classifier in the training sample. The curve

was smoothed for better visualization. *(C)* The tensorboard monitor graph of sex classifier in the validation sample.

**Performance of the AD classifier**

After attaining an industrial-grade brain imaging-based classifier for sex with high cross-dataset accuracy, we used transfer learning to see if we could classify patients with AD. The structure of the model for sex classification was transferred, and the weights of the last two layers (e.g., full connection layer and drop out layer) were reset. The model was fine-tuned on the ADNI sample (4715 images from normal controls (NC) and 2206 images from participants with AD). To ensure generalizability, we utilized five-fold cross-site validation to assess classifier accuracy. Of note, images from same person or same sites were not allocated into both training and testing samples, i.e., every image in a testing sample was collected from participants and sites that were not in the training sample. The AD classifier achieved an average accuracy of 88.4% (accuracy = 92.1%, 82.8%, 88.5%, 90.9% and 85.3% in 5 folds) in the test samples. Average sensitivity and specificity are 0.814 and 0.917, respectively. The ROC AUC reached 0.938 when results from the 5 testing samples were taken together (see Fig. 3).

To test the generalizability of the AD classifier, we tested it on an unseen independent AD dataset, i.e., Open Access Series of Imaging Studies 1/2 (22, 23) (OASIS 1/2; adopting the same criteria as ADNI, 1272 image samples met ADNI criteria for AD or NC; AD/NC = 277/995). The AD classifier achieved 86.1% accuracy in OASIS with 0.921 AUC (see Fig. S1*A*). Sensitivity and specificity were 0.789 and 0.881, respectively. To ensure specificity to AD, we also tested it on MDD, ASD and ADHD samples. The model achieved 86.4% accuracy (e.g., only 13.6% of samples in MDD, ASD and ADHD were misclassified as AD patients; 94.2%, 77.1% and 81.4% accuracy for REST-meta-MDD, ABIDE1/2 and ADHD200 samples) in this test, yielding specificity comparable to that for the OASIS sample, indicates high specificity of this AD classifier in diverse patient samples.

Importantly, although the AD classifier is agnostic to mild cognitive impairment (MCI), we directly tested it on the MCI dataset in ADNI to see if it has the potential to predict the progression of MCI to AD. For the MCI patients who eventually converted to AD, the classifier predicted 63.2% as AD. For those MCI patients who did not convert to AD during the ADNI data collection period, only 22.1% were classified as AD (see Fig. S1*B*). Those results suggested that the classifier is practical for screening MCI patients who have a higher risk of progression to AD. In sum, we believe our AD classifier can provide important insights relevant to computer aided diagnosis and prediction of AD, thus we have freely provided it on the website http://brainimagenet.org:8088. Importantly, classification results by the online classifier should be interpreted with caution, as they cannot replace the diagnosis of licensed clinicians.

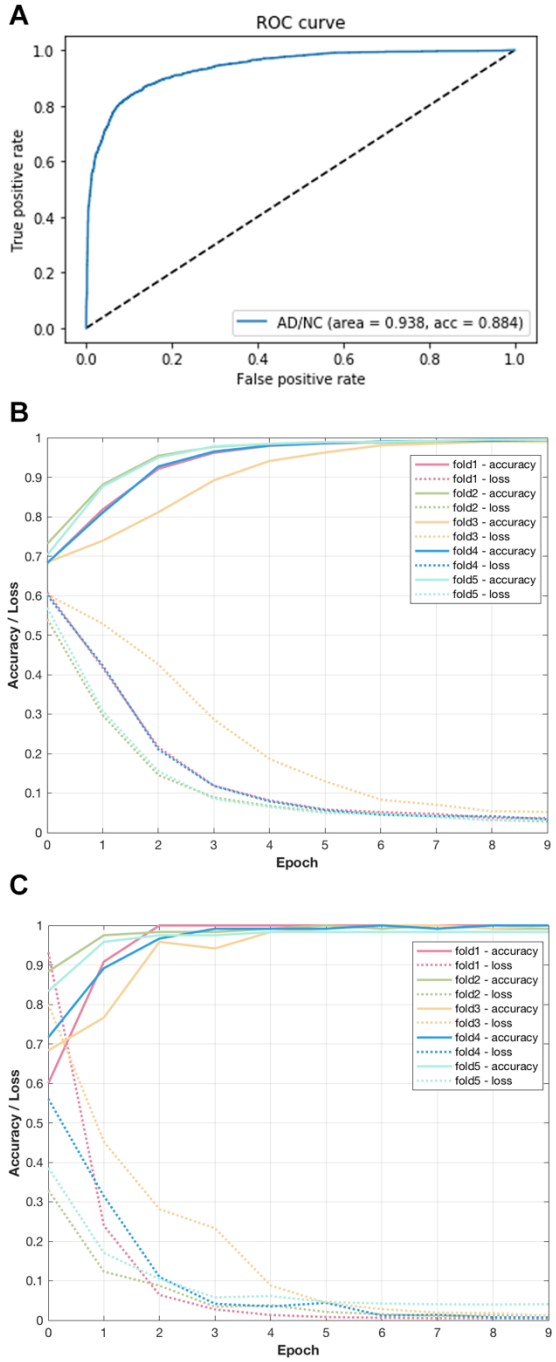


**Fig. 3.** Performance of the Alzheimer's disease (AD) classifier. *(A)* The receiver operating characteristic curve of the AD classifier. *(B)* The tensorboard monitor panel of the AD classifier in the training sample. *(C)* The tensorboard monitor panel of the AD classifier in the validation sample.

**Performance of the classifiers for psychiatric disorders**

We further applied this transfer learning framework to MDD, ASD and ADHD samples to see if it would work for these psychiatric disorders. The training and testing procedures were the same as those for the AD classifier. To ensure generalizability, we utilized five-fold cross-site validation to

assess classifier accuracy. The MDD/NC classifier achieved 55.6% accuracy in the testing sample with AUC of 0.562. The ADHD classifier achieved 63.1% accuracy with AUC of 0.669. The ASD classifier achieved 57% accuracy with AUC of 0.604 (see Fig. S2-S4, left panel). Importantly, the performances of our cross-site validations were all worse than those of traditional random cross-validation (69.3% accuracy for the MDD classifier, 69.4% accuracy for the ADHD classifier, 58.3% accuracy for the ASD classifier, see Fig. S2-S4, right panel). This indicates that classifiers for psychiatric disorders are more sensitive to site variability, thus a useful model should be fine-tuned for each specific site.

**Interpretation of the deep learning classifiers**

To better understand the brain imaging-based deep learning classifier, we calculated occlusion maps for the classifiers. In brief, we continuously set a cubic brain area of every input image to zeros, and made the classifier attempt classification based on the defective samples. The accuracy achieved on "intact" samples by the classifier minus accuracy achieved on "defective" samples indicated the "importance" of the blanked brain area for the classifier. The occlusion map showed that hypothalamus, superior vermis, thalamus, amygdala, putamen, accumbens, hippocampus and parahippocampal gyrus played critical roles in predicting sex (see Fig. 4*A*). The occlusion map for the AD classifier highlighted that hippocampus, parahippocampal gyrus, putamen and insula also played unique roles in predicting AD (see Fig. 4*B*).

To investigate the clinical significance of the output of the AD classifier, we calculated the Spearman's correlation coefficient between the predicted scores by the classifier and mini-mental state examination (MMSE) scores in AD, NC and MCI samples. This analysis confirmed significant negative correlations between the predicted scores and MMSE scores for AD ($r$ = -0.3186, $p < 1 \times 10^{-40}$), NC ($r$ = -0.1086, $p < 1 \times 10^{-10}$), MCI ($r$ = -0.4079, $p < 1 \times 10^{-188}$) and the overall samples ($r$ = -0.5787, $p < 1 \times 10^{-188}$) (See Fig. 5). As lower MMSE scores indicated more severe of cognitive impair for AD and MCI patients, we confirmed that the more severe the disease, the higher predicted score by the classifier. In addition, both the predicted scores and MMSE scores showed significant differences between MCI patients with further conversion and MCI patients without further conversion (predicted scores: $t$ = 13.4545, $p < 1\times 10^{-36}$, Cohen's d = 1.0298; MMSE scores: $t$ = -8.0146, $p < 1\times 10^{-14}$, Cohen's d = -0.6134). Importantly, the effect size of the predicted scores by the classifier is much larger than the behavioral measure (MMSE scores).

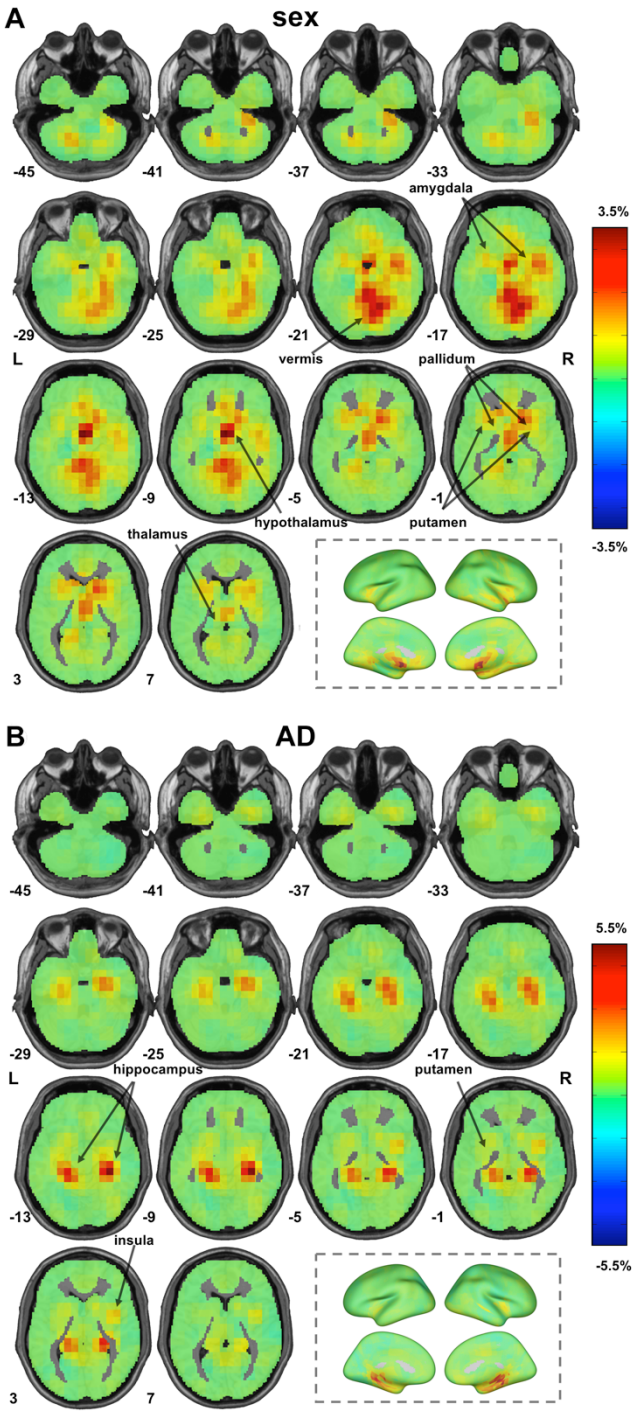


**Fig. 4.** Interpretation of the deep learning classifiers with occlusion maps. Classifier performance dropped considerably when the brain areas rendered in red were masked out of the model input.

*(A)* The occlusion maps for the sex classifier. *(B)* The occlusion maps for Alzheimer's disease classifier. The graphs on the bottom right show the occlusion maps projected to the brain surface.

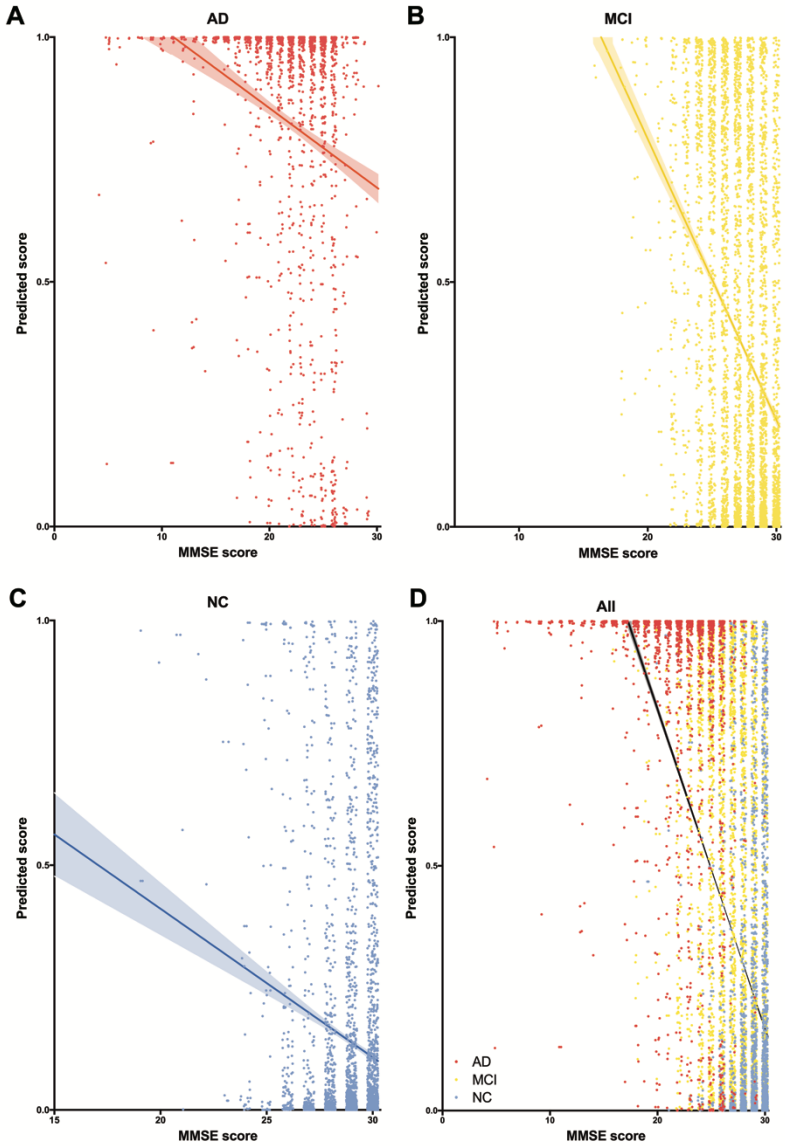


**Fig. 5.** Correlations between the output of Alzheimer's disease (AD) classifier and the severity of illness. The predicted scores from AD classifier showed significant negative correlations with the mini-mental state examination (MMSE) scores of AD, normal control (NC) and mild cognitive impairment (MCI) samples. *(A)* Correlation between the predicted scores from AD classifier and the MMSE scores of AD samples. *(B)* Correlation between the predicted scores from AD classifier and the MMSE scores of NC samples. *(C)* Correlation between the predicted scores from AD classifier and the MMSE scores of MCI samples. *(D)* Correlation between the predicted scores from AD classifier and the MMSE scores of AD, NC and MCI samples.

## Discussion

Using an unprecedentedly large sample, we built an industrial-grade classifier for sex which can classify the sex of a participant with brain structural imaging data from anyone and any scanner

with about 95% accuracy. Using transfer learning, the model fine-tuned to AD achieved 88.4% accuracy in stringent cross-validation and 86.1% accuracy for a direct test on an unseen independent dataset (OASIS). Predicted scores of the AD classifier showed significant negative correlations with the severity of illness ($r$ = -0.5787). When we directly tested the AD classifier on brain images of MCI patients, 63.2% of those who eventually converted to AD were predicted as AD, versus 22.1% of those who did not convert to AD during the ADNI data collection period were predicted as AD. The AD classifier also achieved high specificity in direct testing on other datasets (e.g., MDD, ADHD, ASD). Occlusion test on this model showed that hypothalamus, superior vermis, thalamus, amygdala and limbic system areas were critical for predicting sex and hippocampus, parahippocampal gyrus, putamen and insula played key roles in predicting AD. Finally, the transfer learning framework failed to achieve useful accuracy for psychiatric disorders, raising issues for future studies.

The industrial grade high accuracy and generalizability (95% for anyone and any scanner) of our deep neural network sex classifier demonstrates brain imaging can have practical utility for predicting individual differences. The current prototype may facilitate future research to apply brain imaging in many practical application fields. Of note, the output of the deep neural network model is a continuous variable, thus the threshold can be adjusted to balance sensitivity and specificity. For example, when tested on the independent sample (OASIS), sensitivity and specificity results were 0.789 and 0.881, respectively, as the default threshold was set at 0.5. However, for screening, the false negative rate should be minimized even at the cost of higher false positive rates. If we lower the threshold (e.g., to 0.3), sensitivity can be improved to 0.893 at a cost of decreasing specificity to 0.773. Thus, in our freely available AD prediction website (http://brainimagenet.org:8088), users can obtain continuous outputs and adjust the threshold by themselves. In addition, the output of the deep neural network model showed significant negative correlations with MMSE scores. Considering the “greedy” characteristic of deep neural network for reducing training loss, the perdition scores for AD and NC may be overstated, and those negative correlations were even underestimated. When direct comparing the predicted scores (or MMSE scores) between MCI subjects with and without further conversion, the predicted scores showed much higher effect size than MMSE scores (Cohen’d $_{prediction}$ = 1.0298 vs Cohen’d $_{MMSE}$ = -0.6134), indicating that predicted scores from AD classifier have better prompting/warning effects for physician to differentiating MCI patients.

Sex differences, especially in the human brain, is an old but lingering question (24, 25). Testosterone was first identified as masculinizing the brain. Since, several steroid hormones like estradiol and progesterone as well as genetic and environmental factors have been shown to affect differences between brains of males and females (24). Daphna and colleagues extracted hundreds of voxel-based morphometry (VBM) features from structural MRI and concluded that “the so-called male/female brain” does not exist as no single structural feature can support a sexually dimorphic view of human brains (26). However, human brains can embody sexually dimorphic features in a multivariate manner. The high accuracy and high generalizability of the sex classifier in the present study demonstrates that sex was separable in a 1,981,440-dimension (96*120*86*2) feature space. Among those 1,981,440 features, occlusion analysis revealed that the features located in hypothalamus played the most critical roles in predicting sex. The hypothalamus regulates testosterone secretion through hypothalamic-pituitary-gonadal axis thus plays a critical role in brain masculinization (27). Men have significantly larger hypothalamus than women relative to cerebrum size (28). Other than the hypothalamus, broad brain areas in the occlusion map including left amygdala, posterior cingulate, left accumbens, parahippocampal gyri and hippocampus also affected classification performance. These regions can also be summarized as pertaining to the limbic system and they support a variety of functions including emotion, motivation and long-term memory. Males and females also show differences in these brain structures (29). Taken together, our machine learning evidence shows that the “male/female brain” does exist.

As with sex differences, AD induced brain structural changes have been frequently reported by structural MRI studies. Among all the structural findings, hippocampus atrophy is the most prominent change and is used in imaging assisted diagnosis (30). Hippocampus (and entorhinal cortex) neurobiological changes precede progressive neocortical damage and AD symptoms (31). The convergence of our deep learning system and human physicians on hippocampus structure transformation for classifying AD patients further supports the crucial role of the hippocampus in AD. Other than hippocampus, differential atrophy has also been observed in putamen (32) and insula (33) in AD patients compared to normally aging adults. We speculate that the lower accuracy of the AD classifier than the sex classifier reflects greater biological heterogeneity in AD.

For psychiatric disorders, our model failed to achieve practical accuracy. Importantly, there are still no objective biomarkers for psychiatric disorders, including MDD, ASD and ADHD. The accuracy and consistency of diagnoses by clinicians are themselves suboptimal (e.g., for diagnosing MDD, sensitivity ranges from 0.25 to 0.95, specificity ranges from 0.33 to 0.95, depending on the instruments used (34)). As psychiatric disorder labels can be inaccurate, any brain image-based classifier trained on these samples cannot yield better accuracy than the input labels (diagnoses by clinicians). Accordingly, we did not expect high accuracy of the model for psychiatric disorders. Future studies utilizing information on prognosis and treatment response will have the potential to transform the diagnosis and treatment of mental disorders. When such data become available, artificial intelligence (AI) systems can improve the efficiency and reliability of the diagnostic process.

In the deep learning field, the appearance of ImageNet (35) tremendously accelerated the evolution of computer vision. As data organization and preprocessing of MRI data require tremendous time, manpower and computational loads, these constraints impede scientists from other fields entering brain imaging. Open access preprocessed brain imaging big data is fundamental to facilitate the participation of a broader range of researchers. Beyond building and sharing an industrial-grade brain imaging-based deep learning classifier, we invite researchers (especially computer scientists) to join the effort to decipher the brain by openly sharing all sharable preprocessed data (Link_To_Be_Added upon publication, preprocessed data of some datasets could not be shared as the raw data owners do not allow sharing data derivatives).

We also openly share our models to allow other researchers to directly deploy them (https://github.com/Chaogan-Yan/BrainImageNet). The training of the 3-dimensional Inception-ResNet-V2 in the present study was powered by 4 NVIDIA Tesla V100 32G GPU. However, researchers do not need to buy expensive GPUs but can instead deploy the compressed model directly on much cheaper computers. Our code is openly shared as well (https://github.com/Chaogan-Yan/BrainImageNet), thus allowing other researchers to replicate the present results and further develop brain imaging-based classifiers based on our existing work. Finally, we have also built a demonstration website for classifying sex and AD (http://brainimagenet.org:8088). Users can upload their own raw T1 or preprocessed GMD and GMV data to acquire the predictions of sex or AD labels in real time.

Study limitations should be acknowledged. Considering the lower reproducibility of functional MRI compared to structural MRI, only structural MRI derived images were used in the present deep learning model. But functional physiology may further improve the performance of sex and brain disorder classifiers. In future studies, functional MRI, especially resting-state functional MRI, has the potential to provide additional information for model training. Furthermore, with advances in software like FreeSurfer (36), fmriprep (37) and DPABISurf, surface-based algorithms may take the place of volume-based algorithms. Surface-based algorithms are more time and computation consuming, but can provide more precise brain registration and reproducibility (38). Future studies should take surface-based images as inputs of deep learning models.

In summary, we pooled 34 openly-shared datasets to constitute the biggest brain MRI sample (n = 85,721) to date, and applied a state-of-the-art architecture deep convolutional neural network,

Inception-ResNet-V2, to build an industrial-grade sex classifier. The AD classifier obtained through transfer learning reached high accuracy and sufficient generalizability to be of practical use, thus demonstrating the feasibility of transfer learning in brain disorder applications. Future work is needed to deploy such a framework in psychiatric disorders and other aspects of individual differences.

## Materials and Methods

### Data acquisition

We submitted data access applications to nearly all the open-access brain imaging data archives, and received permissions from the administrators of 34 datasets. The full dataset list is shown in Table S1. Deidentified data were contributed from datasets approved by local Institutional Review Boards. The reanalysis on these data were approved by the Institutional Review Board of Institute of Psychology, Chinese Academy of Sciences. All study participants provided written informed consent at their local institution. All 50,876 participants (contributing 85,721 samples) had at least one session with a T1-weighted structural image and information on sex and age. For participants with multiple sessions of structural images, each image was considered as an independent sample for data augmentation in training. Importantly, scans from the same person were never split into training and testing sets, as that would artifactually inflate performance. To test if our classifier could be transferred to brain/psychiatric disorders, we selected ADNI (n = 16,590), REST-meta-MDD (n = 2,380), Autism Brain Imaging Data Exchange (ABIDE) 1&2 (n = 2,145) and ADHD200 (n = 875) datasets.

### MRI preprocessing

We did not feed raw data to classifier training, but used the knowledge in brain imaging data analysis. The brain structural data were segmented and normalized to acquire grey matter density (GMD) and grey matter volume (GMV) maps. Specifically, the voxel based morphometry (VBM) analysis module within Data Processing Assistant for Resting-State fMRI (39) (DPARSF), which was developed based on SPM (40), was used to segment individual T1-weighted images into GM, WM and cerebrospinal fluid (CSF). Then, the segmented images were transformed from individual native space to MNI space (a coordinate system created by Montreal Neurological Institute) by the Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (41) (DARTEL) tool. Two voxel-based structural metrics, GMD and GMV were fed into the deep learning classifier as two features for each participant. GMV was modulated GMD images by using the Jacobian determinants derived from the spatial normalization in the VBM analysis (42).

### Quality control

Bad quality raw structural images would produce distorted GMD and GMV maps during segmentation and normalization. To remove such participants from affecting the training classifiers, we excluded participants in each dataset with spatial correlation exceeding the threshold defined by mean - 2SD Pearson's correlation between each participant's GMV map and the grand mean GMV template. The grand mean GMV template was generated by randomly selecting 10 participants from each dataset and averaging the GMV maps of all these 340 (from 34 datasets) participants. All these participants were visually checked for imaging quality. After quality control, 83,735 samples were retained for classifier training.

### Deep learning: classifier training and testing for sex

We trained a 3-dimension Inception-ResNet-v2 (21) model adopted from its 2-dimension version in the Keras built-in application (see Fig. 1*A* for structure). This is a record-breaking model in pattern recognition (21) which integrates two classical series of CNN models, Inception and ResNet. We replaced the convolution, pooling and normalization modules with their 3-dimension versions and adjusted the number of layers and convolutional kernels to make them suitable for 3-dimension MRI inputs (e.g., GMD and GMV as different input channels). The present model consists of one

stem module, three groups of convolutional modules (Inception-ResNet-A/B/C) and two reduction modules (Reduction-A/B). It can take advantage of convolutional kernels with different sizes and shapes and extract features in different sizes, and mitigate vanishing gradients and exploding gradients by adding residual modules. We utilized the Keras built-in stochastic gradient descent optimizer with learning rate = 0.01, nesterov momentum = 0.9, decay = 0.001 (e.g., learn rate = learn rate0 x (1 / (1 + decay x batch))). Loss function was set to binary cross-entropy. The batch size was set to 24 and the training procedure lasted 10 epochs for each fold. To avoid potential overfitting, we randomly split 600 samples out of the training sample as a validating sample and set a checking point at the end of every epoch. We saved the model in which the epoch classifier showed the lowest validating loss. Thereafter, the testing sample was fed into this model to test the classifier.

To ensure generalizability, we used cross-dataset validation. In the testing phase, all the data from a given dataset would never be seen during the classifier training phase. This also ensured the data from a given site (and thus a given scanner) were unseen by the classifier during training. This strict setting can limit classifier performance, however, this made it feasible to generalize to any participant at any site (scanner). Five-fold cross-dataset validation was used to assess classifier accuracy. Of note, 3 datasets were always kept in the training sample due to the massive number of samples: Adolescent Brain Cognition Development (ABCD) (n = 31,176), UK Biobank (n = 20,124) and Alzheimer's Disease Neuroimaging Initiative (ADNI) (n = 16,596). The remaining 31 datasets were randomly allocated into the training and testing samples. The allocating schemas were the solution that balanced the sample size of 5 folds the best from 10,000 random allocating procedures.

**Transfer learning: classifier training for AD**

After obtaining an industrial-grade brain imaging-based classifier for sex with high cross-dataset accuracy, we used transfer learning to see if we could classify AD patients. The structure of the sex model was kept, and weights in the last two layers (e.g., full connection layer and drop out layer) were reset. This new model was transferred to the ADNI dataset (n = 16,596, adni.loni.usc.edu). ADNI was launched in 2003 (Principal Investigator Michael W. Weiner, MD) to investigate biological markers of the progression of MCI and early AD (see www.adni-info.org). Five-fold cross-site validation was used to assess classifier accuracy. By ensuring the data from a given site (and thus a given scanner) were unseen by the classifier during training, this strict strategy made the classifier generalizable with non-inflated accuracy, thus better simulating realistic medical applications than traditional five-fold cross-validation.

Furthermore, to test the generalizability of the AD classifier, we directly tested the classifier on another unseen independent AD sample, i.e., Open Access Series of Imaging Studies (OASIS) (22, 23). We used averaged output of 5 AD classifiers in previous five-fold cross-validation as the final output for a participant. As OASIS did not specify the criteria for an AD diagnosis, we adopted 2 criteria from ADNI-1 to define AD patients, i.e., 1) mini-mental state examination score between 20 and 26 (inclusive) and 2) clinical dementia rating score = 0.5 or 1.0. Thus, we tested on 277 AD samples and 995 normal control samples who met the ADNI-1 criteria of AD and normal controls. Of note, OASIS scanning conditions and recurrent criteria differed much more than variations among different ADNI sites, thus we expected to achieve lower performance. This AD classifier was also tested on MDD, ASD and ADHD samples to see if it had specificity in a more complex sample, i.e., if the patients with mental disorder would be mis-classified as AD patients.

We further investigated whether the AD classifier can predict progression of people with MCI. MCI is a diagnosis defined as cognitive decline without impairment in everyday activities (43). The amnestic subtype of MCI has a high risk of converting to AD. We screened image records of the MCI patients who converted to AD later in ADNI 1/2/go phases, and collected 1668 images labeled as ‘MCI’ (of note, images labeled ‘Conversion: MCI to AD’ and images labeled as ‘AD’ after conversion are not included). 4069 images labeled ‘MCI’ without later conversion were also

assembled for contrast. We directly fed all these MCI images into the AD classifier without further fine-tuning, thus evaluating the performance of the AD classifier on unseen MCI information.

**Transfer learning: classifier training for psychiatric disorders**

We further applied this transfer learning framework on MDD, ASD and ADHD samples to see if it would work for psychiatric disorders. The sex classifier was transferred to three mental disorder samples including REST-meta-MDD, ABIDE 1&2 and ADHD200. The training parameters were the same as for training the AD classifier. After fine tuning, five-fold cross-site validation was used to assess classifier accuracy.

**Interpretation of the deep learning classifier**

To better understand the brain imaging based deep learning classifier, we calculated occlusion maps for the classifiers. We repeated testing the images in testing sample using the model with the highest accuracy within the 5 folds, while successively masking brain areas (volume = 12mm*12mm*12mm, step = 6mm) of all input images. The accuracy achieved on "intact" samples by the classifier minus accuracy achieved on "defective" samples indicated the "importance" of the blanked brain area for the classifier. The occlusion maps were calculated for both sex and AD classifiers. To investigate the clinical significance of the output of the AD classifier, we calculated the Spearman's correlation coefficient between the predicted scores and mini-mental state examination (MMSE) scores of AD, NC and MCI samples. For verifying whether the predicted scores (or MMSE score) have difference between MCI patients with further conversion to AD and MCI patients without further conversion to AD, we carried out two-sample t-test on the predicted scores (or MMSE score) between MCI patients with further conversion and MCI patients without further conversion. We randomly selected one T1 image scan for each MCI subject and finally collected 234 MCI patients with further conversion and 631 MCI patients without further conversion.

**Data and Code availability**

The imaging, phenotype and clinical data used for the training, validation and test sets were applied from the administrators of 34 datasets. The preprocessed brain imaging data would be available on (Link_To_Be_Added upon publication), some data can be available after acquiring the permission of administrators of original datasets. The code for training and testing the model are openly shared at https://github.com/Chaogan-Yan/BrainImageNet.

## Acknowledgments

The authors appreciate the editorial assistance and support of Dr. Francisco Xavier Castellanos. This work was supported by the National Key R&D Program of China (grant number: 2017YFC1309902), the National Natural Science Foundation of China (grant number: 81671774, 81630031), the 13th Five-year Informatization Plan of Chinese Academy of Sciences (grant number: XXH13505), the Key Research Program of the Chinese Academy of Sciences (grant NO. ZDBS-SSW-JSC006), Beijing Nova Program of Science and Technology (grant number: Z191100001119104). Data used in the preparation of this article for training and testing the sex classifier was obtained from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9-10 and follow them over 10 years into early adulthood. The ABCD Study is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/scientists/workgroups/. ABCD consortium investigators designed and

implemented the study and/or provided data but did not necessarily participate in analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. This research has been conducted using the UK Biobank Resource. Data collection and sharing for the training and testing the sex and AD classifier were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

**References**

1. Zhang C, Dougherty CC, Baum SA, White T, & Michael AM (2018) Functional connectivity predicts gender: Evidence for gender differences in resting brain connectivity. *Hum. Brain Mapp.* 39(4):1765-1776.
2. Luo ZG, Hou CP, Wang LB, & Hu DW (2019) Gender Identification of Human Cortical 3-D Morphology Using Hierarchical Sparsity. *Front. Hum. Neurosci.* 13:29.
3. Kaufmann T*, et al.* (2019) Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nat. Neurosci.* 22(10):1617-1623.
4. Jonsson BA*, et al.* (2019) Brain age prediction using deep learning uncovers associated sequence variants. *Nat Commun* 10(1):5409.
5. Kong R*, et al.* (2019) Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality, and Emotion. *Cereb. Cortex* 29(6):2533-2551.
6. Perrin RJ, Fagan AM, & Holtzman DM (2009) Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. *Nature* 461(7266):916-922.
7. Challis E*, et al.* (2015) Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *NeuroImage* 112:232-243.
8. Drysdale AT*, et al.* (2017) Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat. Med.* 23(1):28-38.
9. Fonzo GA*, et al.* (2019) Brain regulation of emotional conflict predicts antidepressant treatment response for depression. *Nat Hum Behav* 3(12):1319-1331.
10. Bellec P*, et al.* (2017) The Neuro Bureau ADHD-200 Preprocessed repository. *NeuroImage* 144(Pt B):275-286.
11. Hazlett HC*, et al.* (2017) Early brain development in infants at high risk for autism spectrum disorder. *Nature* 542(7641):348-351.

12. Emerson RW*, et al.* (2017) Functional neuroimaging of high-risk 6-month-old infants predicts a diagnosis of autism at 24 months of age. *Sci. Transl. Med.* 9(393).
13. Ham Y-G, Kim J-H, & Luo J-J (2019) Deep learning for multi-year ENSO forecasts. *Nature* 573(7775):568-572.
14. DeVries PMR, Viegas F, Wattenberg M, & Meade BJ (2018) Deep learning of aftershock patterns following large earthquakes. *Nature* 560(7720):632-634.
15. Liu WB*, et al.* (2017) A survey of deep neural network architectures and their applications. *Neurocomputing* 234:11-26.
16. Kermany DS*, et al.* (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172(5):1122-1131.
17. Esteva A*, et al.* (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115-118.
18. McKinney SM*, et al.* (2020) International evaluation of an AI system for breast cancer screening. *Nature* 577(7788):89-94.
19. He T*, et al.* (2020) Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage* 206:116276.
20. Pan SJ & Yang Q (2010) A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10):1345-1359.
21. Szegedy C, Ioffe S, Vanhoucke V, & Alemi AA (2016) Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *National Conference on Artificial Intelligence*, pp 4278-4284.
22. Marcus DS*, et al.* (2007) Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.* 19(9):1498-1507.
23. Marcus DS, Fotenos AF, Csernansky JG, Morris JC, & Buckner RL (2010) Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *J. Cogn. Neurosci.* 22(12):2677-2684.
24. Arnold AP (2012) The end of gonad-centric sex determination in mammals. *Trends Genet.* 28(2):55-61.
25. Shaywitz BA*, et al.* (1995) Sex differences in the functional organization of the brain for language. *Nature* 373(6515):607-609.
26. Joel D*, et al.* (2015) Sex beyond the genitalia: The human brain mosaic. *Proc. Natl. Acad. Sci. U. S. A.* 112(50):15468-15473.
27. Forest MG, Peretti ED, & Bertrand J (1976) Hypothalamic-pituitary-gonadal relationships in man from birth to puberty. *Clin. Endocrinol. (Oxf.)* 5(5):551-569.
28. Makris N*, et al.* (2013) Volumetric parcellation methodology of the human hypothalamus in neuroimaging: Normative data and sex differences. *NeuroImage* 69:1-10.
29. Becker JB (1999) Gender differences in dopaminergic function in striatum and nucleus accumbens. *Pharmacol. Biochem. Behav.* 64(4):803-812.
30. Frisoni GB, Fox NC, Jack CR, Jr., Scheltens P, & Thompson PM (2010) The clinical use of structural MRI in Alzheimer disease. *Nat. Rev. Neurol.* 6(2):67-77.
31. Braak H & Braak E (1991) Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* 82(4):239-259.

32. de Jong LW*, et al.* (2008) Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. *Brain* 131(12):3277-3285.
33. Rombouts SA, Barkhof F, Witter MP, & Scheltens P (2000) Unbiased whole-brain analysis of gray matter loss in Alzheimer's disease. *Neurosci. Lett.* 285(3):231-233.
34. Pettersson A, Bostrom KB, Gustavsson P, & Ekselius L (2015) Which instruments to support diagnosis of depression have sufficient accuracy? A systematic review. *Nord J Psychiatry* 69(7):497-508.
35. Deng J*, et al.* (2009) Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, (Ieee), pp 248-255.
36. Fischl B (2012) FreeSurfer. *NeuroImage* 62(2):774-781.
37. Esteban O*, et al.* (2019) fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Med.* 16(1):111-116.
38. Coalson TS, Van Essen DC, & Glasser MF (2018) The impact of traditional neuroimaging methods on the spatial localization of cortical areas. *Proc. Natl. Acad. Sci. U. S. A.* 115(27):e6356-e6365.
39. Yan CG & Zang YF (2010) DPARSF: A MATLAB Toolbox for "Pipeline" Data Analysis of Resting-State fMRI. *Front. Syst. Neurosci.* 4:13.
40. Friston KJ*, et al.* (1994) Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* 2(4):189-210.
41. Goto M*, et al.* (2013) Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra provides reduced effect of scanner for cortex volumetry with atlas-based method in healthy subjects. *Neuroradiology* 55(7):869-875.
42. Good CD*, et al.* (2001) A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage* 14(1 Pt 1):21-36.
43. Gauthier S*, et al.* (2006) Mild cognitive impairment. *The lancet* 367(9518):1262-1270.

## Figures and Tables

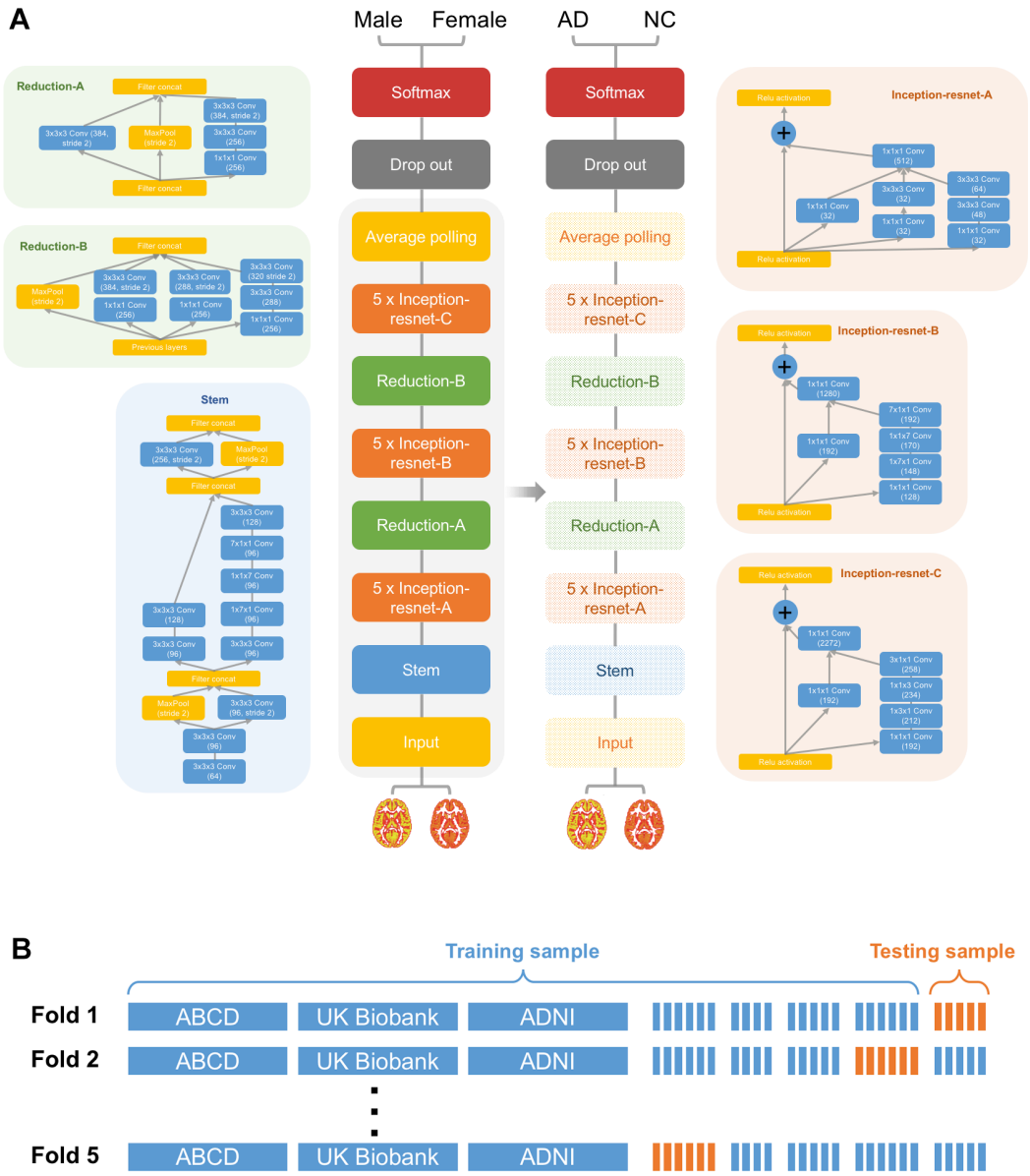


**Figure 1.** Flow diagram for the training procedure for the sex classifier and transfer learning framework. *(A)* Schema for 3D Inception-ResNet-V2 network and transfer learning framework. *(B)* Schematic diagram for leave-dataset-out 5-fold cross-validation in training the sex classifier.

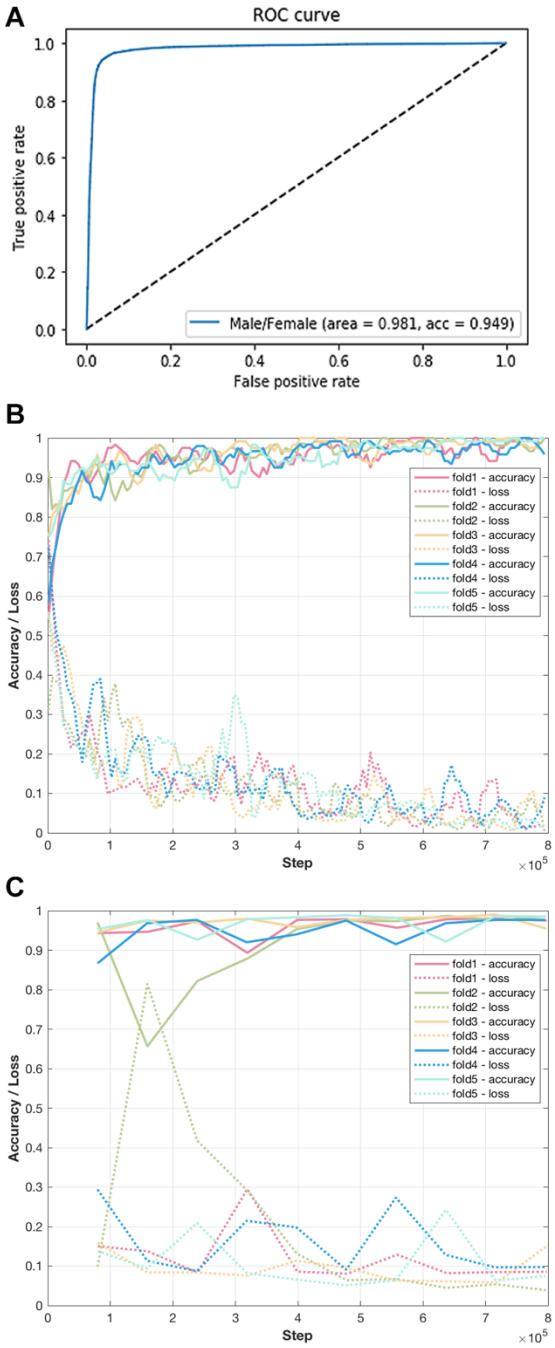


**Figure 2.** Performance of the sex classifier. *(A)* The receiver operating characteristic curve of the sex classifier. *(B)* The tensorboard monitor graph of the sex classifier in the training sample. The curve was smoothed for better visualization. *(C)* The tensorboard monitor graph of sex classifier in the validation sample.

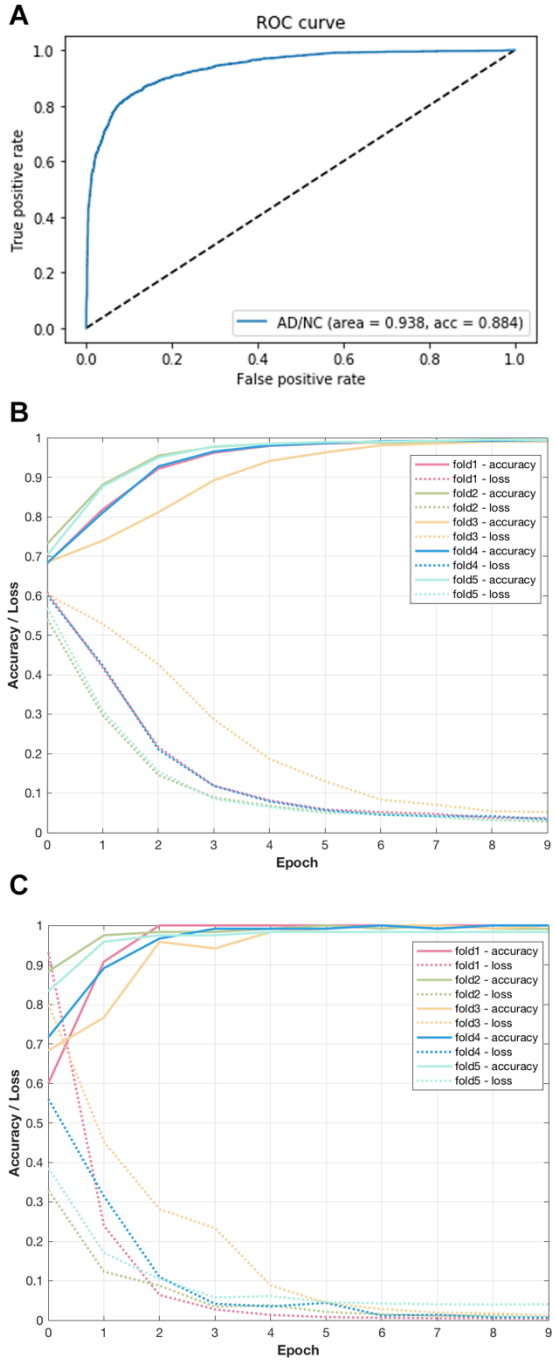


**Figure 3.** Performance of the Alzheimer's disease (AD) classifier. *(A)* The receiver operating characteristic curve of the AD classifier. *(B)* The tensorboard monitor panel of the AD classifier in the training sample. *(C)* The tensorboard monitor panel of the AD classifier in the validation sample.

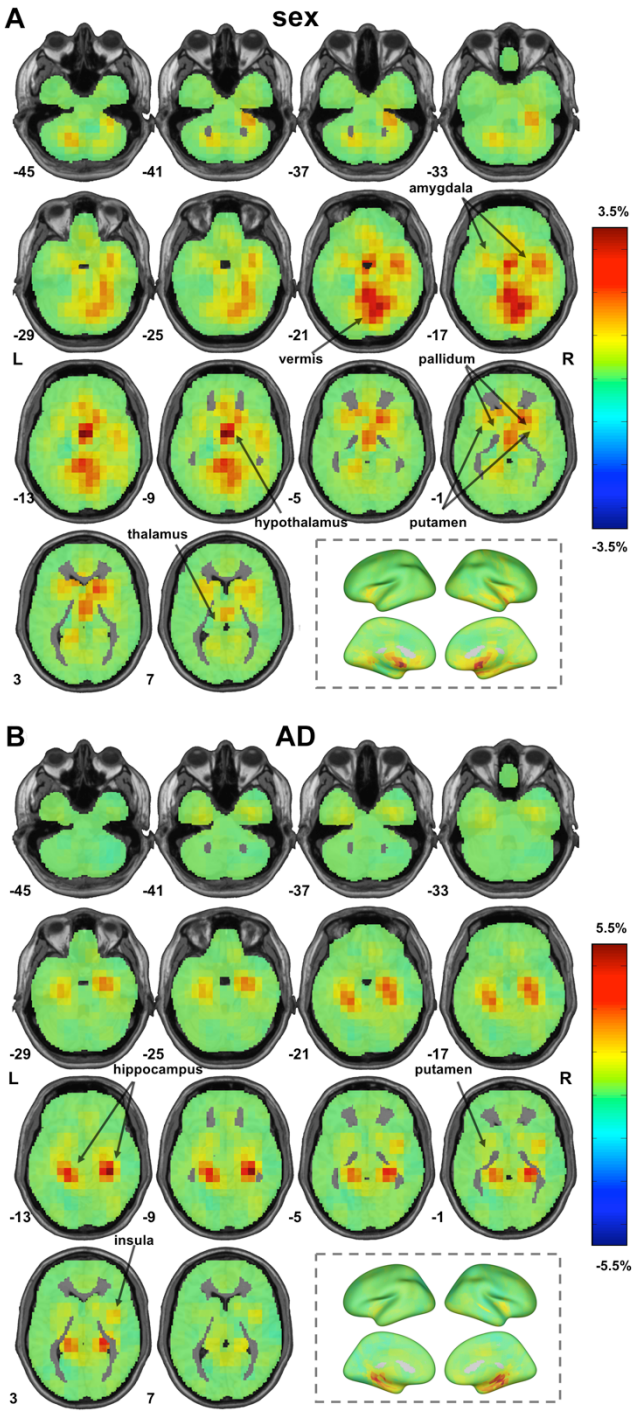


**Figure 4.** Interpretation of the deep learning classifiers with occlusion maps. Classifier performance dropped considerably when the brain areas rendered in red were masked out of the model input. *(A)* The occlusion maps for the sex classifier. *(B)* The occlusion maps for Alzheimer disease classifier. The graphs on the bottom right show the occlusion maps projected to the brain surface.

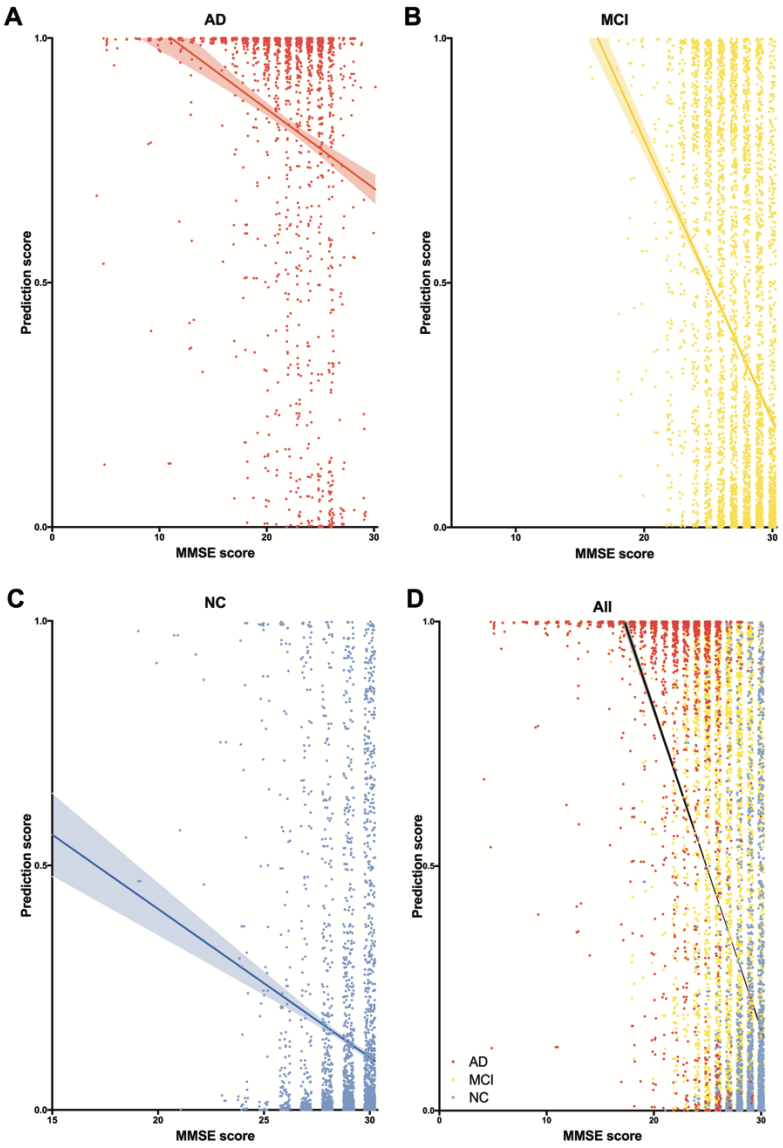


**Fig. 5.** Correlations between the output of Alzheimer's disease (AD) classifier and the severity of illness. The predicted scores from AD classifier showed significant negative correlations with the mini-mental state examination (MMSE) scores of AD, normal control (NC) and mild cognitive impairment (MCI) samples. *(A)* Correlation between the predicted scores from AD classifier and the MMSE scores of AD samples. *(B)* Correlation between the predicted scores from AD classifier and the MMSE scores of NC samples. *(C)* Correlation between the predicted scores from AD classifier and the MMSE scores of MCI samples. *(D)* Correlation between the predicted scores from AD classifier and the MMSE scores of AD, NC and MCI samples.