Classification of Sex and Alzheimer's Disease via Brain Imaging-Based Deep Learning on 85,721 Samples

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. **Part of the 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

Running title

Brain Deep Learning model for Sex and AD

Abstract

Beyond detecting brain damage or tumors, little success has been attained on identifying individual differences and brain disorders with magnetic resonance imaging (MRI). Here, we sought to build industrial-grade brain imaging-based classifiers to infer two types of such inter-individual differences: sex and Alzheimer’s disease (AD), using deep learning/transfer learning on big data. We pooled brain structural data from 217 sites/scanners to constitute the largest brain MRI sample to date (85,721 samples from 50,876 participants), and applied a state-of-the-art deep convolutional neural network, Inception-ResNet-V2, to build a sex classifier with high generalizability. In cross-dataset-validation, the sex classification model was able to classify the sex of any participant with brain structural imaging data from any scanner with 94.9% accuracy. We then applied transfer learning based on this model to objectively diagnose AD, achieving 88.4% accuracy in cross-site-validation on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and 91.2% / 86.1% accuracy for a direct test on two unseen independent datasets (AIBL / OASIS). Directly testing this AD classifier on brain images of unseen mild cognitive impairment (MCI) patients, the model correctly predicted 63.2% who eventually converted into AD, versus predicting 22.1% as AD who did not convert into AD during follow-up. Predicted scores of the AD classifier correlated significantly with illness severity. By contrast, the transfer learning framework was unable to achieve practical accuracy for psychiatric disorders. To improve interpretability of the deep learning models, occlusion tests revealed that hypothalamus, superior vermis, thalamus, amygdala and limbic system areas were critical for predicting sex; hippocampus, parahippocampal gyrus, putamen and insula played key roles in predicting AD. Our trained model, code, preprocessed data and an online prediction website have been openly-shared to advance the clinical utility of brain imaging.

Keywords

Alzheimer’s disease, brain MRI, convolutional neural network, sex difference, transfer learning

1. Introduction

Can we infer individual differences and brain disorders from brain images? This is a question that has been long pursued. However, beyond visually identifying brain damage or tumors, little success has been attained in identifying individual differences, e.g., age, sex, or brain disorders, e.g., Alzheimer's disease (AD). These may contain subtle features that cannot be discerned by visual inspection, but which may be amenable to identification based on machine intelligence. Here, we sought to build industrial-grade brain imaging-based classifiers for sex and AD with high generalizability via deep learning/transfer learning on big data.

Progress has been attained in using brain imaging, especially magnetic resonance imaging (MRI), to predict sex,[1,2] age,[3,4] Alzheimer's Disease (AD),[5,6] major depressive disorder (MDD),[7,8] attention-deficit/hyperactivity disorder (ADHD),[9] and autism spectrum disorder (ASD) among others.[10,11] However, all of these efforts have failed to generalize. Brain imaging data varies depending on characteristics such as scanner vendor, head coil type, imaging sequence, applied gradient fields, reconstruction methods, 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,[12] aftershock pattern prediction[13] and automatic speech recognition.[14] In medical imaging, image-based deep convolutional neural networks (CNN) have been applied to objectively diagnose retinal diseases,[15] skin cancer[16] and breast cancer screening.[17] 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.

Taking AD diagnosis as an example, safe and non-invasive MRI-based biomarkers are needed to supplement current invasive diagnostic biomarkers like cerebrospinal fluid (CSF), amyloid positron emission tomography (PET) and tau imaging.[18-20] However, prior attempts have yet to reach clinical utility. Qiu and colleagues[21] built an interpretable deep-learning classifier for AD with an average accuracy of 82.2% using brain imaging data from four datasets. However, the performance of the proposed AD classifier is quite unstable across datasets. For example, in AIBL dataset, the AD classifier achieved 87.0% accuracy and 0.924 specificity but with a deficient 0.594 sensitivity. On the contrary, in FHS dataset, the accuracy of the same classifier dropped to 76.6% with high sensitivity (0.901) and inadequate specificity (0.712). The floating accuracy and inconsistent tradeoff between sensitivity and specificity in different medical units hampered the proposed method to be integrated into the present diagnosis system. To alleviate the unsatisfactory generalization performance, Bashyam et al.[22] used a more heterogeneous sample to build a brain age prediction model that would be more generalizable to unseen sites/scanners. However, when transfer learning to AD, they only used random cross-validation on the ADNI dataset with an accuracy of 86% and didn't implement independent dataset validation. Random cross-validation may share participants from the same sites between training and testing samples, thus the model may not apply to datasets from unseen sites due to the site information leaking in training. To attain generalizability, cross-dataset or cross-site validation should be implemented to make sure classifier accuracy will be insensitive to site/scanner variability.

A bottleneck for developing an industrial-grade brain imaging-based classifier is the needed 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 (45,615 participants) has come mainly from a single site (UKBiobank).[3] The second and third largest training data sets comprised 16,848 and 14,468 participants.[4,22] Even with a relatively large sample, if the training sample doesn't contain sufficient sites (i.e., with variations in manufacturers of MR equipment, scanning parameters, quality control procedures and participant characteristics, etc.), a trained classifier will fail to generalize to unseen datasets. Thus, here, we utilized the largest and most diversiform sample to date (85,721 samples from 50,876 participants from 217 sites/scanners, see Table S1), to achieve an industrial-grade classifier which can generalize to any scanner and any sample.

Our first training goal was to predict sex, as it is an objective dichotomous indicator available for every participant in open datasets. After obtaining a brain imaging-based classifier for sex with high cross-dataset accuracy, our second goal was to use transfer learning to attempt to classify patients with AD. Transfer learning is preferred as the AD dataset is much smaller, and direct training on a small sample can result in overfitting with poor generalization to new unseen testing data [23]. The third goal was to test the specificity of our AD model on MDD, ASD and ADHD datasets, and to explore the transfer learning framework to these psychiatric disorders. This study advanced brain imaging-based deep-learning towards clinical utilities in four ways. First, we implemented big data on an unprecedented scale, comprising 85,721 samples from 217 sites/scanners, thus permitting us to build an industrial-grade brain imaging-based deep learning classifier. Second, as generalizability is crucial for practical use, we always used stringent cross-dataset-validation or cross-site-validation during training/testing, thus allowing our model to be generalized to anybody from any site/scanner. Third, other than the traditional 2D Inception-ResNet-v2 deep neural network models, the 3D neural network we expanded reflects the 3D nature of the brain and improves interpretability through occlusion testing. Lastly, we openly shared our preprocessed data, trained model, code and framework to facilitate open-science, and have built an online prediction website (http://brainimagenet.org:8088) for anyone interested in testing our classifier with brain imaging data from anybody and any scanner.

2. 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. Reanalyses of these data were approved by the Institutional Review Board of the 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 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 disorders, we selected ADNI (16,596 samples from 2,212 participants), Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL, 624 samples from 406 participants), Open Access Series of Imaging Studies (OASIS, 3,150 samples from 1,664 participants), REST-meta-MDD (2,380 participants), Autism Brain Imaging Data Exchange (ABIDE) 1&2 (2,145 participants) and ADHD200 (875 participants) datasets.

MRI preprocessing

We did not feed raw data for classifier training, but used the knowledge from brain imaging data analysis. 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 (DPARSF),[24] which was developed based on SPM,[25] 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) with the Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DARTEL) tool.[26] The two voxel-based structural metrics, GMD and GMV, were fed into the deep learning classifier as two channels per participant. GMV was modulated GMD images using the Jacobian determinants derived from the spatial normalization in the VBM analysis.[27]

Quality control

Poor quality raw structural images produce distorted GMD and GMV maps during segmentation and normalization. To prevent such participants from affecting the training classifiers, we excluded participants in each dataset with spatial correlation lower than 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. The image quality of all 340 scans was visually checked. After quality control, 83,735 samples from 49,558 participants were retained for classifier training.

Deep learning: classifier training and testing for sex

We trained a 3-dimension Inception-ResNet-v2 model adopted from its 2-dimension version in the Keras built-in application (see Fig. 1A for structure).[28] This is a record-breaking model in pattern recognition 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. 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 on the data of 83,735 samples from 49,558 participants with 34 datasets scanned from 217 sites/scanners. 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. While this strict setting inevitably limits classifier performance, 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 after quality control: Adolescent Brain Cognition Development (ABCD) (30,533 samples from 11,875 participants), UK Biobank (19,760 participants) and Alzheimer's Disease Neuroimaging Initiative (ADNI) (16,431 samples from 2,212 participants). The remaining 31 datasets were randomly allocated to 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.

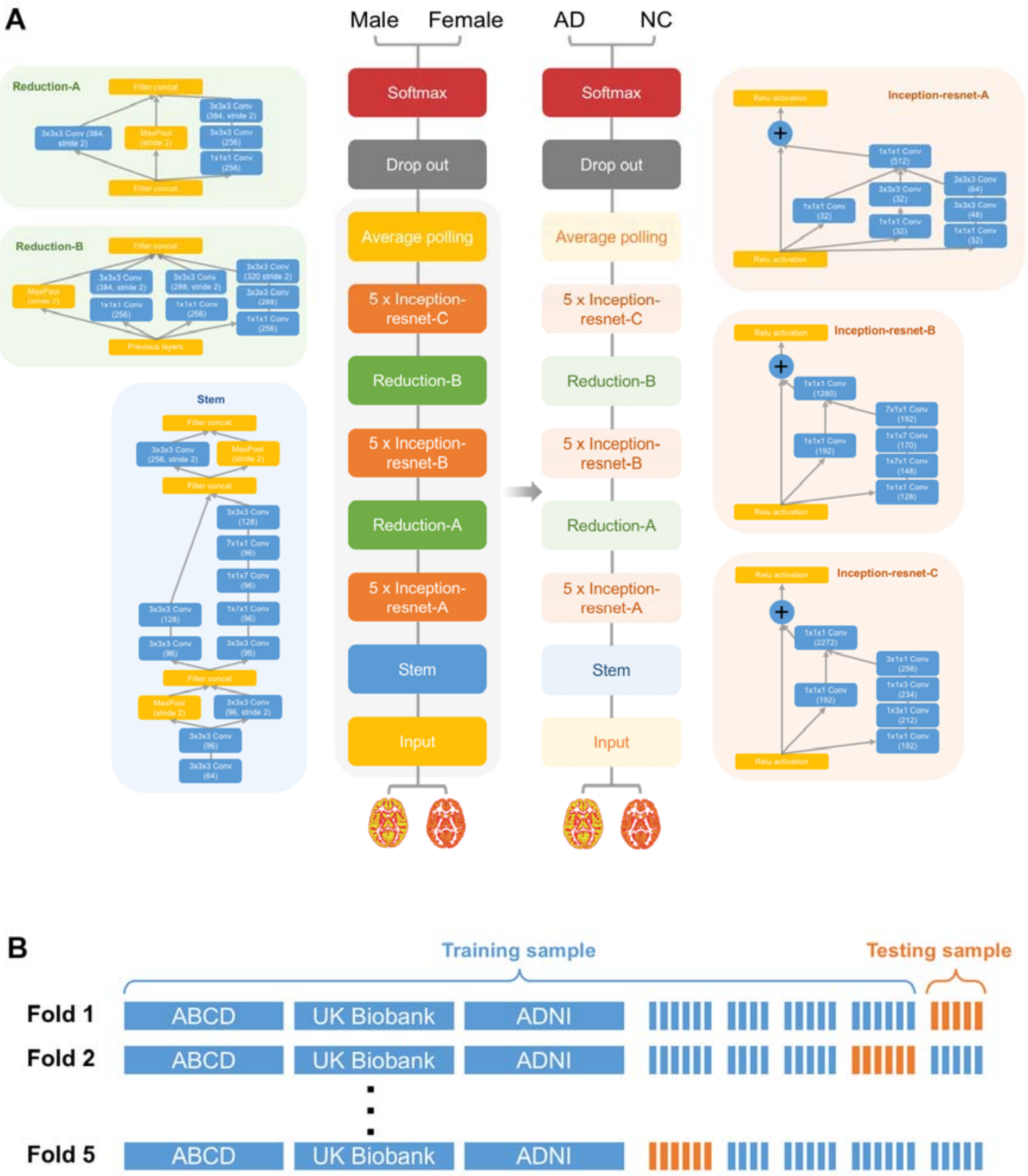

Figure 1 | Flow diagram for training procedure for the sex classifier and the Alzheimer's disease transfer learning framework. (**A**) Schema for 3D Inception-ResNet-V2 network and the Alzheimer's disease transfer learning framework. (**B**) Schematic diagram for leave-dataset-out 5-fold cross-validation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

Transfer learning: classifier training and testing 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 (2,186 samples from 380 AD patients and 4,671 samples from 698 normal controls (NCs)). 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 random cross-validation.

To further test the generalizability of the AD classifier, we directly tested the classifier on two unseen independent AD sample, i.e., AIBL[29] and OASIS[30,31]. We used the averaged output of 5 AD classifiers in the previous five-fold cross-site-validation as the final output for a participant. We used diagnoses provided by AIBL dataset as the labels of samples (101 samples from 82 AD patients and 523 samples from 324 NCs). 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 patients and 995 NCs who met the ADNI-1 criteria of AD and NCs in OASIS dataset. Of note, AIBL and 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 determine its specificity in a more complex sample, i.e., would patients with mental disorders be misclassified as AD patients.

We further investigated whether the AD classifier could predict the progression of MCI. MCI is defined as cognitive decline without impairment in everyday activities.[32] The amnestic subtype of MCI has a high risk of converting to AD. We screened image records of the MCI patients who subsequently converted to AD in ADNI 1/2/go phases, and collected 1668 samples from 235 participants labeled as 'MCI' (i.e., they had follow-up visits labeled as

'Conversion: MCI to AD' or 'AD', but images acquired at those follow-up visits were not used). We also assembled 4069 samples from 624 participants labeled 'MCI' without later conversion for contrast. We fed all these MCI images directly into the AD classifier without further fine-tuning, thus evaluating the performance of the AD classifier on unseen MCI information.

Transfer learning: classifier training and testing for psychiatric disorders

We further applied this transfer learning framework to MDD, ASD and ADHD samples to determine its performance with psychiatric disorders. The sex classifier was transferred to psychiatric samples from REST-meta-MDD (1266 MDDs vs. 1097 NCs), ABIDE 1&2 (985 ASDs vs. 1107 NCs) and ADHD200 (181 ADHDs vs. 526 NCs) after quality control. The training parameters were the same used for training the AD classifier. After fine-tuning, five-fold cross-site-validation was used to assess classifier accuracy.

Interpretation of the deep learning classifiers

To better understand the brain imaging-based deep learning classifier, we calculated occlusion maps for the classifiers. We repeatedly tested images in the testing sample using the model with the highest five-fold accuracy, while successively masking brain areas (volume = 18mm*18mm*18mm, step = 9mm) in all input images. The accuracy achieved with "intact" samples by the classifier minus accuracy achieved with "defective" samples indicated the "importance" of the occluded brain area for the classifier. Occlusion maps were calculated for both sex and AD classifiers.

Data and code availability

The imaging, phenotype and clinical data used for the training, validation and test sets were obtained from the administrators of 34 datasets. The raw data are publicly available in different repositories. The preprocessed brain imaging data are available through the R-fMRI Maps project (Link_To_Be_Added upon publication; preprocessed data for some datasets could not be shared as the raw data owners do not allow sharing data derivatives). The code for training and testing the model are openly shared at https://github.com/Chaogan-Yan/BrainImageNet. The online prediction website is available at http://brainimagenet.org:8088.

3. Results

Brain imaging big data

Only brain imaging data with sufficient size and variety can make deep learning useful for building an industrial-grade classifier. We received permissions from the administrators of 34 datasets (85,721 samples from 50,876 participants from 217 sites/scanners, see Table S1; some datasets did not require application). 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 adapted from its 2-dimension version in the Keras built-in application (see Fig. 1A for structure). To ensure generalizability, five-fold cross-dataset-validation was used to assess classifier accuracy (see Fig. 1B). The five-fold cross-dataset-validation accuracies were: 94.8%, 94.0%, 94.8%, 95.7% and 95.8%, for an overall average accuracy of 94.9% in testing samples. Area under the 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 about 95% accuracy. Interested readers can test this model at the online prediction website (http://brainimagenet.org:8088). The code and model are also openly shared at https://github.com/Chaogan-Yan/BrainImageNet.

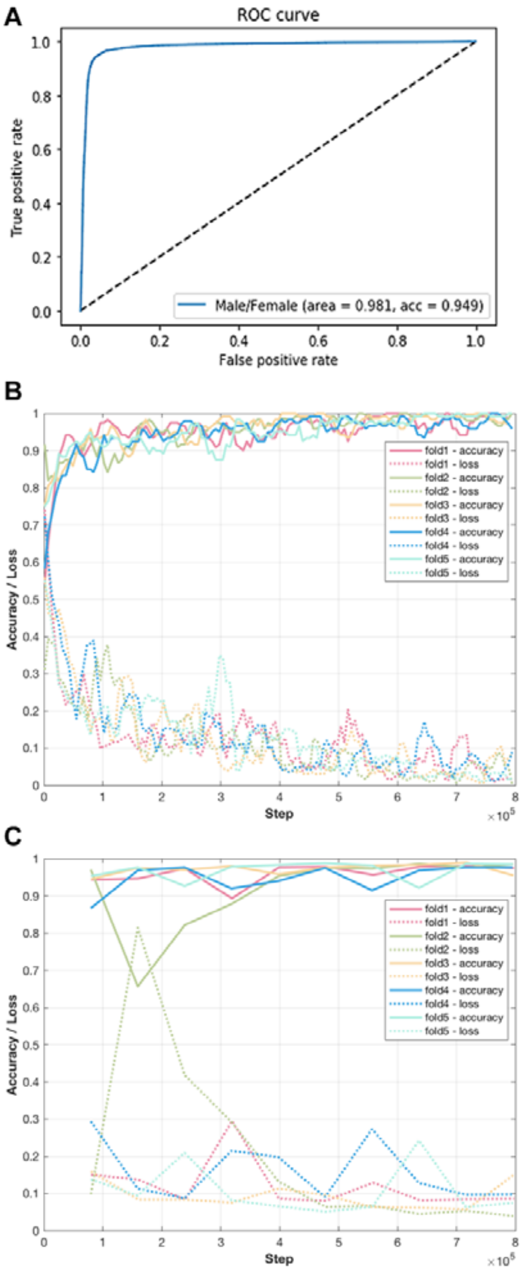

Figure 2 | Performance of the sex classifier. (**A**) Receiver operating characteristic curve of the sex classifier. (**B**) Tensorboard monitor graph of the sex classifier in the training sample. The curve was smoothed for better visualization. (**C**) Tensorboard monitor graph of sex classifier in 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. To ensure generalizability, we utilized five-fold cross-site-validation to assess classifier accuracy. 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 ADNI test samples. Average sensitivity and specificity were 0.814 and 0.917, respectively. The ROC AUC reached 0.938 when results from the 5 testing samples were combined (see Fig. 3 and Table. 1).

To test the generalizability of the AD classifier, we applied it to an unseen independent AD dataset, i.e., AIBL and OASIS 1/2. The AD classifier achieved 91.2% accuracy in AIBL with 0.948 AUC (see Table. 1 and Supplementary Fig. 1A). Sensitivity and specificity were 0.851 and 0.924, respectively. The AD classifier achieved 86.1% accuracy in OASIS with 0.921 AUC (see Table. 1 and Supplementary Fig. 1B). Sensitivity and specificity were 0.789 and 0.881, respectively. To assess 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 MDD, ASD or ADHD samples were misclassified as AD; 94.2%, 77.1% and 81.4% accuracy for REST-meta-MDD, ABIDE1/2 and ADHD200 samples, respectively) in this test, yielding specificity comparable to that for the OASIS sample, indicating high specificity of this AD classifier in diverse patient samples.

Table 1 | performance of the Alzheimer's disease classifier

Dataset	n (AD)	n (NC)	Accuracy	AUC	Sensitivity	Specificity
ADNI	2186	4671	0.884	0.938	0.814	0.917
AIBL	101	523	0.912	0.948	0.851	0.924
OASIS	277	995	0.861	0.921	0.789	0.881

AD = Alzheimer's disease; NC = normal control. The sample sizes showed here the numbers of T1 MRI scans.

Importantly, although the AD classifier is agnostic to mild cognitive impairment (MCI), we directly tested it on the MCI dataset in ADNI to determine its potential to predict progression from MCI to AD. For MCI patients who eventually converted to AD, the classifier predicted

63.2% as AD. For MCI patients who did not convert to AD during the ADNI data collection period, only 22.1% were classified as AD (see Supplementary Fig. 1C). These results suggest that the classifier is practical for screening MCI patients with a higher risk of progression to AD. In sum, we believe our AD classifier can support computer-aided diagnosis and prediction of AD, thus we have made it freely available at http://brainimagenet.org:8088. Nevertheless, we emphasize that online classification results should be interpreted with caution, as they cannot replace evaluation and diagnosis by licensed clinicians.

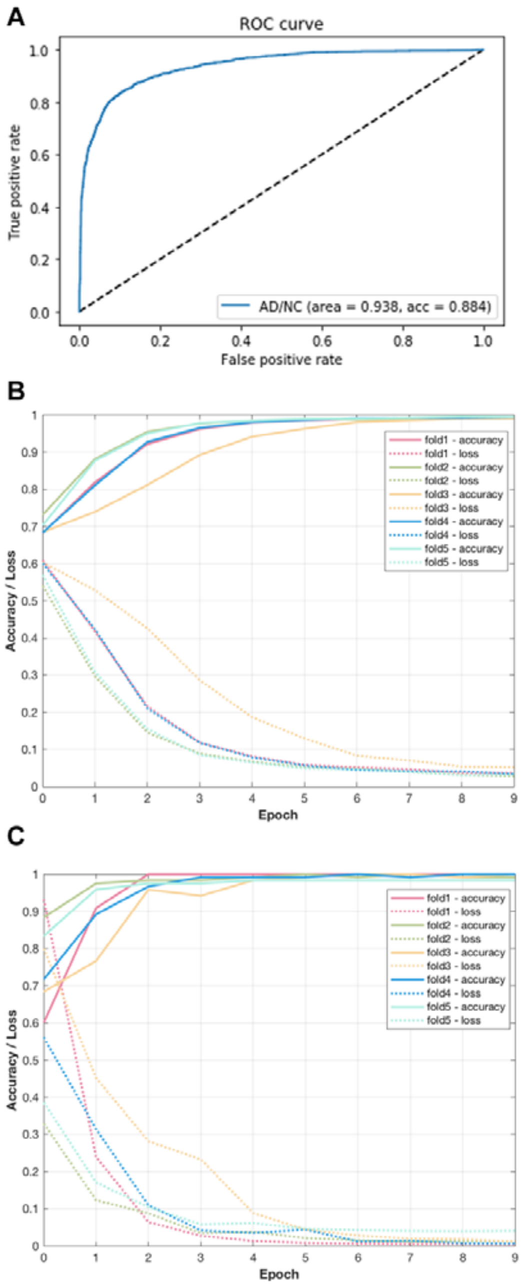

Figure 3 | Performance of the Alzheimer's disease (AD) classifier. (**A**) Receiver operating characteristic curve of the AD classifier. (**B**) Tensorboard monitor panel of the AD classifier in the training sample. (**C**) Tensorboard monitor panel of the AD classifier in the validation sample.

Performance of the classifiers for psychiatric disorders

We also applied this transfer learning framework to MDD, ASD and ADHD samples to determine its performance 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 Supplementary Figs. 2-4, left panel). Notably, the performance 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 Supplementary Figs. 2-4, 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 attempted classification with the defective samples. Occlusion maps showed that hypothalamus, superior vermis, thalamus, amygdala, putamen, accumbens, hippocampus and parahippocampal gyrus played critical roles in predicting sex (see Fig. 4A). Occlusion maps for the AD classifier highlighted hippocampus, parahippocampal gyrus, putamen and insula as playing unique roles in predicting AD (see Fig. 4B).

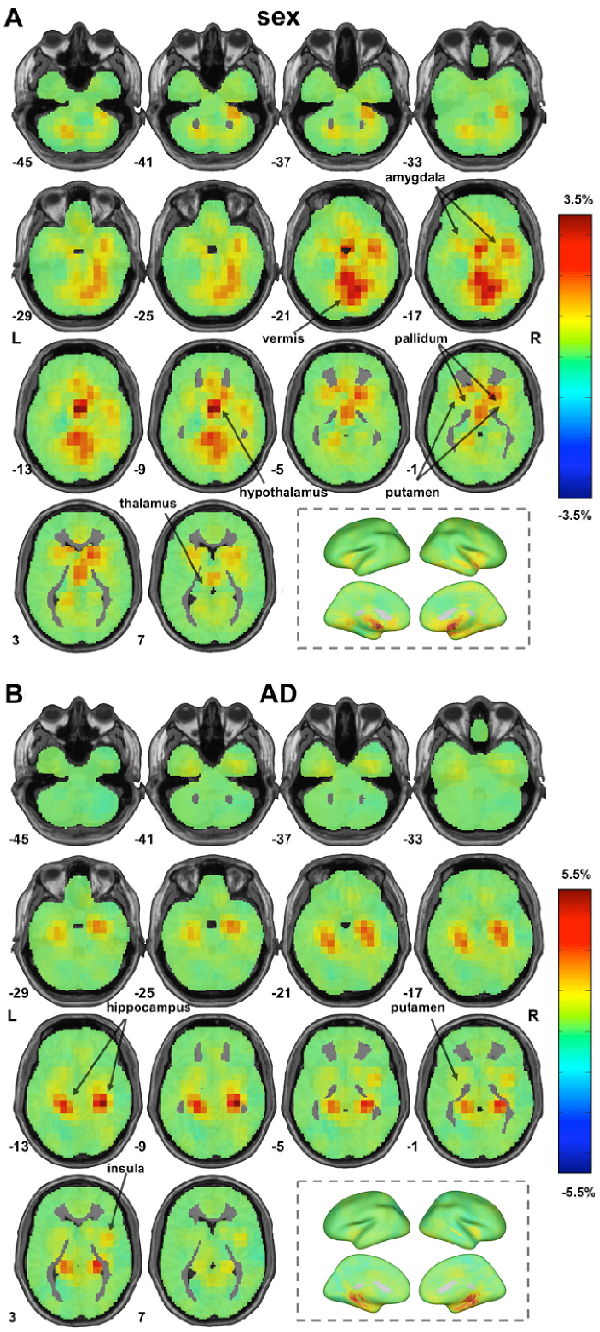

Figure 4 | Interpretation of the deep learning classifiers with occlusion maps. Classifier performance dropped considerably when the brain areas rendered in red were masked out of the model input. (**A**) Occlusion maps for the sex classifier. (**B**) Occlusion maps for the Alzheimer disease classifier.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

To investigate the clinical significance of the output of the AD classifier, we calculated Spearman's correlation coefficients between scores predicted by the classifier and mini-mental state examination (MMSE) scores in AD, NC and MCI samples. We observed significant negative correlations between predicted scores and MMSE scores for AD (r = -0.319, $p < 1 \times 10^{-40}$), NC (r = -0.109, $p < 1 \times 10^{-10}$), MCI (r = -0.408, $p < 1 \times 10^{-188}$) and the overall sample (r = -0.579, $p < 1 \times 10^{-188}$) (See Fig. 5). As lower MMSE scores indicated more severe cognitive impairment for AD and MCI patients, we confirmed that the more severe the disease, the higher the predicted score by the classifier. In addition, both predicted scores and MMSE scores differed significantly between MCI patients who converted to AD and those who did not (predicted scores: t = 13.454, $p < 1\times 10^{-36}$, Cohen's d = 1.03; MMSE scores: t = -8.015, $p < 1\times 10^{-14}$, Cohen's d = -0.61) (See Supplementary Fig. 5). Importantly, the effect size of the scores predicted by the classifier is much larger than the behavioral measure (MMSE scores).

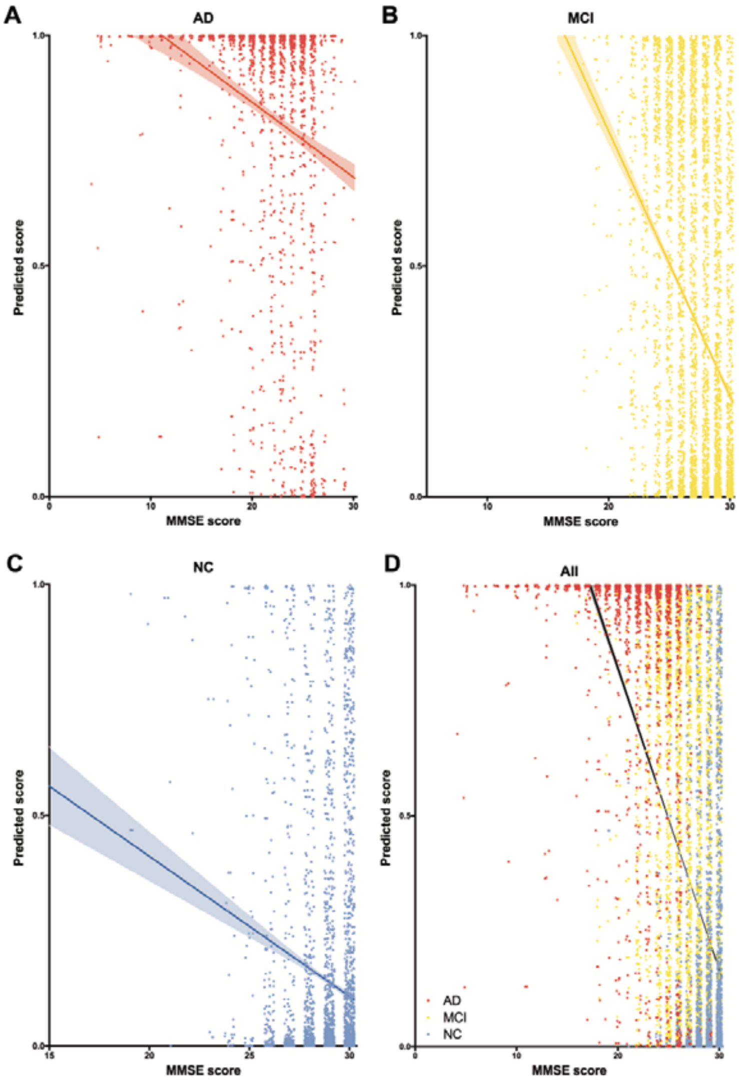

Figure 5 | Correlations between Alzheimer's disease (AD) classifier output and illness severity. The scores predicted by the AD classifier were significantly negatively correlated with mini-mental state examination (MMSE) scores of AD, normal control (NC) and mild cognitive impairment (MCI) samples. (**A**) Correlations between scores predicted by the AD classifier and MMSE scores of AD samples. (**B**) Correlations between scores predicted by the AD classifier and MMSE scores of NC samples. (**C**) Correlations between scores predicted by the AD classifier and MMSE scores of MCI samples. (**D**) Correlations between scores predicted by the AD classifier and MMSE scores of AD, NC and MCI samples.

4. 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-site-validation and 91.2% / 86.1% accuracy for direct tests on unseen independent dataset (AIBL and OASIS). Predicted scores of the AD classifier were significantly negatively correlated with illness severity (*r* = -0.579). When we directly tested the AD classifier on brain images of unseen 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 follow-up interval. The AD classifier also achieved high specificity in direct testing on other datasets (e.g., MDD, ADHD, ASD). Occlusion tests 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. By contrast, the transfer learning framework failed to achieve useful accuracy for psychiatric disorders.

The industrial-grade accuracy and generalizability (95% and 88% for sex and AD, respectively, for anyone and any scanner) of our deep neural network classifiers demonstrates brain imaging can have practical utility for predicting individual differences (e.g., sex and AD). The current prototype should be amenable to other brain imaging applications. The deep neural network model output is a continuous variable; thus, the threshold can be adjusted to balance sensitivity and specificity. For example, when testing the AD model on the independent sample (OASIS), sensitivity and specificity results were 0.789 and 0.881, respectively, when 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 openly available AD prediction website (http://brainimagenet.org:8088), users can obtain continuous outputs and adjust the threshold by themselves. This adjustable characteristic of the model makes it feasible to integrate it into diagnostic criteria as a potential diagnostic MRI biomarker. The relatively high sensitivity of our proposed MRI-based biomarker addresses the lower sensitivity of current criteria (even with invasive CSF and PET examinations, sensitivities of IWG-1 and NIA-AA criteria have been reported to be 68% and 65.6%, respectively).[33,34]

Beyond the feasibility of being integrated into diagnostic criteria, the presented AD model also showed outstanding characteristics to be a progression biomarker. First, the output of the deep neural network model was significantly negatively correlated with MMSE scores, although they were not included in model training. Considering the "greedy" characteristic of deep neural networks for reducing training loss, the predicted scores for AD and NC may be overstated, and the magnitude of the negative correlations may have been underestimated. Second, the present model can quantify disease milestones by predicting the progression of MCI patients. MCI patients who eventually converted to AD were more than twice as likely to be predicted as AD than MCI patients who did not convert (63.2% vs 23.1%). Third, when directly comparing predicted scores (or MMSE scores) between MCI subjects with and without conversion to AD, the effect size for predicted scores was much higher than for MMSE scores ($d_{prediction}$ = 1.03 vs. d_{MMSE} = -0.61), indicating that the AD classifier predicted scores provide better prompting/warning effects for physicians seeking to differentiate MCI patients.

Although deep-learning algorithms are described as "black boxes" for their weak interpretability, occlusion analyses showed that the current MRI-based AD biomarker was aligned with published pathological findings and clinical experience. For example, AD induced brain structural changes have been frequently reported in structural MRI studies, with the most prominent change of hippocampus atrophy being used in imaging assisted diagnosis.[35] Hippocampus (and entorhinal cortex) neurobiological changes precede progressive neocortical damage and AD symptoms.[36] 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 the hippocampus, differential atrophy has also been observed in putamen and insula in AD patients compared to normal aging adults.[37,38] We speculate that the lower accuracy of the AD classifier than the sex classifier reflects greater biological heterogeneity in AD, as non-AD dementias (such as vascular dementia, frontotemporal degeneration, dementia with Lewy bodies) may confound AD diagnosis.[35]

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 clinician diagnoses 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).[39] As psychiatric disorder labels can be inaccurate, any brain image-based classifier trained on these samples cannot yield better accuracy than the input labels (clinician diagnoses). Accordingly, we did not expect high model accuracies for psychiatric disorders. Future studies utilizing longitudinal information on prognosis and treatment response would have the potential to transform the diagnosis and treatment of mental disorders. When such data become available, we believe artificial intelligence systems will improve the efficiency and reliability of the diagnostic process.

Our base model can precisely predict the sex of a given participant, thus advancing our understanding of sex differences in the human brain. 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.[40] 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 demonstrated 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 features located in hypothalamus played the most critical role in predicting sex. The hypothalamus regulates testosterone secretion through hypothalamic-pituitary-gonadal axis, thus playing a critical role in brain masculinization.[41] Men have significantly larger hypothalamus than women relative to cerebrum size.[42] Taken together, our machine learning evidence shows that robust “male/female brain” differences do exist.

In the deep learning field, the appearance of ImageNet tremendously accelerated the evolution of computer vision.[43] 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 are 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). Training of the 3-dimensional Inception-ResNet-V2 in the present study was powered by 4 NVIDIA Tesla V100 32G GPUs. However, researchers do not need to buy expensive GPUs but can instead deploy the compressed model directly on much cheaper computers. Our code is also openly shared (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 work to date. Finally, we have built an online prediction website for classifying sex and AD (http://brainimagenet.org:8088). Users can upload their own raw T1 or preprocessed GMD and GMV data to obtain 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. Nevertheless, functional physiology should further improve the performance of sex and brain disorder classifiers. Future studies should examine whether functional MRI, especially resting-state functional MRI, can provide additional information for model training. Furthermore, with advances in software such as FreeSurfer,[44] fmriprep[45] and DPABISurf, surface-based algorithms may replace volume-based algorithms. Surface-based algorithms are more time and computation consuming, but can provide more precise brain registration and reproducibility.[46] Future studies should take surface-based images as inputs of deep learning models. In addition, the present AD classification model was built based on the labels provided by the ADNI database. Future work should incorporate gold standard post-mortem pathological results for AD or treatment response for psychiatric disorders to further advance the clinical value of MRI-based biomarkers.

In summary, we pooled MRI data from 217 sites/scanners to constitute the largest brain MRI sample (85,721 samples) 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 attained high accuracy and sufficient generalizability to be of practical use, thus demonstrating the feasibility of transfer learning in brain disorder applications. Further work is needed to deploy such a framework in psychiatric disorders and other aspects of individual differences.

Acknowledgement

The authors appreciate the editorial assistance and support of Dr. Francisco Xavier Castellanos. 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.

Funding

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).

Competing interests

The authors declare no competing interests.

Supplementary material

Supplementary material is available at Brain online.

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.

Figure legends

Figure 1 Flow diagram for training procedure for the sex classifier and the Alzheimer's disease transfer learning framework. (**A**) Schema for 3D Inception-ResNet-V2 network and the Alzheimer's disease 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**) Receiver operating characteristic curve of the sex classifier. (**B**) Tensorboard monitor graph of the sex classifier in the training sample. The curve was smoothed for better visualization. (**C**) Tensorboard monitor graph of sex classifier in the validation sample.

Figure 3 Performance of the Alzheimer's disease (AD) classifier. (**A**) Receiver operating characteristic curve of the AD classifier. (**B**) Tensorboard monitor panel of the AD classifier in the training sample. (**C**) 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**) Occlusion maps for the sex classifier. (**B**) Occlusion maps for the Alzheimer disease classifier. Graphs on the bottom right show occlusion maps projected to the brain surface.

Figure 5 Correlations between Alzheimer's disease (AD) classifier output and illness severity. The scores predicted by the AD classifier were significantly negatively correlated with mini-mental state examination (MMSE) scores of AD, normal control

(NC) and mild cognitive impairment (MCI) samples. (**A**) Correlations between scores predicted by the AD classifier and MMSE scores of AD samples. (**B**) Correlations between scores predicted by the AD classifier and MMSE scores of NC samples. (**C**) Correlations between scores predicted by the AD classifier and MMSE scores of MCI samples. (**D**) Correlations between scores predicted by the AD classifier and MMSE scores of AD, NC and MCI samples.

References

1 Zhang, C., Dougherty, C. C., Baum, S. A., White, T. & Michael, A. M. Functional connectivity predicts gender: Evidence for gender differences in resting brain connectivity. *Hum. Brain Mapp.* **39**, 1765-1776, (2018).

2 Luo, Z. G., Hou, C. P., Wang, L. B. & Hu, D. W. Gender Identification of Human Cortical 3-D Morphology Using Hierarchical Sparsity. *Front. Hum. Neurosci.* **13**, 29, (2019).

3 Kaufmann, T. *et al.* Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nat. Neurosci.* **22**, 1617-1623, (2019).

4 Jonsson, B. A. *et al.* Brain age prediction using deep learning uncovers associated sequence variants. *Nat Commun* **10**, 5409, (2019).

5 Perrin, R. J., Fagan, A. M. & Holtzman, D. M. Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. *Nature* **461**, 916-922, (2009).

6 Challis, E. *et al.* Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *NeuroImage* **112**, 232-243, (2015).

7 Drysdale, A. T. *et al.* Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat. Med.* **23**, 28-38, (2017).

8 Fonzo, G. A. *et al.* Brain regulation of emotional conflict predicts antidepressant treatment response for depression. *Nat Hum Behav* **3**, 1319-1331, (2019).

9 Bellec, P. *et al.* The Neuro Bureau ADHD-200 Preprocessed repository. *NeuroImage* **144**, 275-286, (2017).

10 Hazlett, H. C. *et al.* Early brain development in infants at high risk for autism spectrum disorder. *Nature* **542**, 348-351, (2017).

11 Emerson, R. W. *et al.* Functional neuroimaging of high-risk 6-month-old infants predicts a diagnosis of autism at 24 months of age. *Sci. Transl. Med.* **9**, (2017).

12 Ham, Y.-G., Kim, J.-H. & Luo, J.-J. Deep learning for multi-year ENSO forecasts. *Nature* **573**, 568-572, (2019).

13 DeVries, P. M. R., Viegas, F., Wattenberg, M. & Meade, B. J. Deep learning of aftershock patterns following large earthquakes. *Nature* **560**, 632-634, (2018).

14 Liu, W. B. *et al.* A survey of deep neural network architectures and their applications. *Neurocomputing* **234**, 11-26, (2017).

15 Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122-1131, (2018).

16 Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**,

115-118, (2017).

17 McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89-94, (2020).

18 Dubois, B. *et al.* Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *The Lancet Neurology* **13**, 614-629, (2014).

19 Jack Jr, C. R. *et al.* Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia* **7**, 257-262, (2011).

20 Dubois, B. *et al.* Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. *The Lancet Neurology* **6**, 734-746, (2007).

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

22 Bashyam, V. M. *et al.* MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain* **143**, 2312-2324, (2020).

23 Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* **22**, 1345-1359, (2010).

24 Yan, C. G. & Zang, Y. F. DPARSF: A MATLAB Toolbox for "Pipeline" Data Analysis of Resting-State fMRI. *Front. Syst. Neurosci.* **4**, 13, (2010).

25 Friston, K. J. *et al.* Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2**, 189-210, (1994).

26 Goto, M. *et al.* Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra provides reduced effect of scanner for cortex volumetry with atlas-based method in healthy subjects. *Neuroradiology* **55**, 869-875, (2013).

27 Good, C. D. *et al.* A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage* **14**, 21-36, (2001).

28 Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. in *National Conference on Artificial Intelligence.* 4278-4284.

29 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. *Alzheimer's & dementia* **6**, 291-296, (2010).

30 Marcus, D. S. *et al.* Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.* **19**, 1498-1507, (2007).

31 Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C. & Buckner, R. L. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *J. Cogn. Neurosci.* **22**, 2677-2684, (2010).

32 Gauthier, S. *et al.* Mild cognitive impairment. *The lancet* **367**, 1262-1270, (2006).

33 de Jager, C. A., Honey, T. E., Birks, J. & Wilcock, G. K. Retrospective evaluation of revised criteria for the diagnosis of Alzheimer's disease using a cohort with post-mortem diagnosis. *Int. J. Geriatr. Psychiatry* **25**, 988-997, (2010).

34 Harris, J. M. *et al.* Do NIA-AA criteria distinguish Alzheimer's disease from frontotemporal dementia? *Alzheimer's & Dementia* **11**, 207-215, (2015).

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

36 Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* **82**,

239-259, (1991).

37 de Jong, L. W. *et al.* Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. *Brain* **131**, 3277-3285, (2008).

38 Rombouts, S. A., Barkhof, F., Witter, M. P. & Scheltens, P. Unbiased whole-brain analysis of gray matter loss in Alzheimer's disease. *Neurosci. Lett.* **285**, 231-233, (2000).

39 Pettersson, A., Bostrom, K. B., Gustavsson, P. & Ekselius, L. Which instruments to support diagnosis of depression have sufficient accuracy? A systematic review. *Nord J Psychiatry* **69**, 497-508, (2015).

40 Joel, D. *et al.* Sex beyond the genitalia: The human brain mosaic. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15468-15473, (2015).

41 Forest, M. G., Peretti, E. D. & Bertrand, J. Hypothalamic-pituitary-gonadal relationships in man from birth to puberty. *Clin. Endocrinol. (Oxf.)* **5**, 551-569, (1976).

42 Makris, N. *et al.* Volumetric parcellation methodology of the human hypothalamus in neuroimaging: Normative data and sex differences. *NeuroImage* **69**, 1-10, (2013).

43 Deng, J. *et al.* in *2009 IEEE conference on computer vision and pattern recognition.* 248-255 (Ieee).

44 Fischl, B. FreeSurfer. *NeuroImage* **62**, 774-781, (2012).

45 Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Med.* **16**, 111-116, (2019).

46 Coalson, T. S., Van Essen, D. C. & Glasser, M. F. The impact of traditional neuroimaging methods on the spatial localization of cortical areas. *Proc. Natl. Acad. Sci. U. S. A.* **115**, e6356-e6365, (2018).