Deep learning system for brain image-aided diagnosis of multiple major mental disorders

Qingfeng Li[1], Wengzheng Wang[1], Qian Guo[1], Lijuan Jiang, [1] Kaini Qiao[1], Yang Hu[1], Xiaochen Zhang[1], Zhen Wang[1], Daihui Peng[1], Qing Fan[1], Min Zhao[1], Yiru Fang[1], Jijun Wang[1], Hong Qiu[1], Jinhong Wang[1], Guanjun Li[1,*], Jianhua Sheng[1,*], Chunbo Li[1], Zhi Yang[1, *], for the Psychiatric Imaging Consortium.

1. Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai 200030, China

Abstract

The current clinical diagnosis of psychiatric disorders relies heavily on subjective assessment of symptoms. While neuroimaging has made an essential contribution to characterizing the brain of psychiatric disorders, it does not currently serve the clinical diagnosis of major psychiatric disorders. Here, we report a neuroimaging-aided diagnostic system for major psychiatric disorders designed for clinical needs. We developed novel deep learning networks with attentional mechanisms and applied them to a large-scale, single-center neuroimaging dataset containing four major psychiatric disorders and healthy groups (n=2490). Both cross-validation and extensive independent validation using multiple open-source datasets (n = 1972) showed that the system could accurately identify any one of the four diagnostic categories and healthy population from brain structural imaging. For the first time, we have constructed an automatic neuroimaging-aid diagnostic system that considers common issues in practice, such as co-morbid diagnoses and the discrimination between specific suspected diagnoses. Furthermore, real-world applications have validated the system's effectiveness. These works contribute to the translation of brain research to objective diagnostic aids for psychiatric disorders.

Introduction

Mental disorders are diseases of the human brain that can cause suffering to the patient or others, severely impair social functioning, and have become the largest economic burden of disease in the world[1]. Bipolar disorder (BD), major depressive disorder (MDD), schizophrenia (SZ), and obsessive-compulsive disorder (OCD) are four clinically common and severe adult psychiatric disorders[2]. Despite significant advances in neurobiological research on these psychiatric disorders, clinical diagnosis

still relies on subjective evaluation of behavioral symptoms[3]. The complexity of symptoms[4-6] and the subjectivity of assessment methods impede accurate and consistent diagnosis. Researchers are striving to find objective techniques to assist in the diagnosis of these psychiatric disorders or subtypes of the disease.

As a non-invasive, high-throughput, and widely available measure, T1-weighted MRI can reflect fine-grained features of brain structure[7-9] that have been associated with environmental and genetic factors for mental disorders[10-20]. Combining machine learning and T1-weighted images to identify certain neurodegenerative diseases, such as Alzheimer's disease and mild cognitive impairment, has seen some success[21-27], supporting the potential to application of this approach as an objective aid for diagnosing major severe psychiatric disorders. Despite its importance for the development of psychiatric diagnosis as a more objective discipline, there is still a lack of neuroimaging-aided diagnostic systems that span a variety of major psychiatric disorders and meet the practical needs of clinical diagnosis.

In clinical practice, the greatest difficulty psychiatrists encounter is not in identifying a diagnostic category in healthy individuals. Instead, they are challenged to identify a person's primary diagnosis from among multiple possible diagnoses, based on a symptom set with considerable individual variation. This is important for the adoption of effective intervention strategies. For example, the depressive episodes of BD may have similar symptoms to those of MDD, but the treatment strategies for the two disorders are different. The wrong treatment strategy may instead exacerbate symptoms. Meanwhile, because symptom overlap and comorbidity are common in the real world[4,6,28], accurate identification of co-morbidity in clinical practice can be difficult.

Current neuroimaging studies do not yet meet these practical needs. Previous attempts have mainly focused on distinguishing certain types of diagnosis from health control[27,29-32]. Some studies have built classifiers to distinguish between two or three diagnoses[33-39], but the sensitivity and specificity of these classifiers have only been tested in two categories and not in the context of multiple diagnostic categories. Neuroimaging-assisted diagnostic systems that infer the possibility of multiple diagnostic types simultaneously and suggest possible co-morbid patterns are lacking.

Furthermore, most classifiers have been validated only using “clean” samples that were collected following strict inclusion and exclusion criteria for scientific research. While the “clean” samples help reduce heterogeneity, they may not reflect the real difficulties in clinical practice. Due to the sophistication of mental illness and the wide range of individual differences, even for an experienced psychiatrist, making a definitive diagnosis at a patent’s first visit is not an easy task. Therefore, whether classifiers trained with scientific data can be used for diagnosis aids in the real world warrant further evaluations.

Here, we construct a neuroimaging-assisted diagnostic system for major psychiatric disorders to meet clinical needs. To achieve this goal, we developed a new spatial patches-based hierarchical deep learning network (PHN) and applied it to a large-scale neuroimaging dataset covering four diagnostic categories and healthy population (n = 2490, SZ, MDD, BD, OCD, HC) to train classifiers to infer multiple

diagnostic types based on T1-weighted MRI brain images. The PHN addresses the methodological challenges through a novel two stage architecture to address the methodological challenges, which applies deep learning methods to localize brain regions of interest and extract image features. Beyond pair-wise classifiers, we obtained classifiers that identify a certain type among multiple diagnostic types. Based on these classifiers, we constructed a closed-loop, automatic neuroimaging-assisted diagnostic system connecting a physician workstation, an MRI workstation, a data analysis server, and a physician query and feedback system, and tested the performance of the system in real-world applications. We found that the PHN classifier can accurately identify primary diagnoses of mental disorders and outperforms common machine learning and deep learning models. Key components of the PHN play essential roles in performance. We validated the performance of the PHN classifier using independent large sample data and verified the ability of the neuroimaging-assisted diagnostic system to identify primary diagnoses and suggest co-morbidities based on consensus diagnoses of psychiatrists in real-world applications.

Results

A patch-based hierarchical network for image classification (PHN)

For accurate recognizing psychiatric diseases, we designed a patch-based hierarchical neural network (PHN) that augmented with a pre-attention stage (PHN-stage1) and local feature extraction and classification stage (PHN-stage2) (Figure 1). PHN-stage1 pre-locates possible disease-related brain regions to reduce redundant information for PHN-stage2. The core engine of PHN is a patch-based feature extraction network (P-FEN), which extracts diagnosis-sensitive features and evaluates disease probability from the input image patch. Unlike the commonly used convolutional neural networks (CNNs), P-FEN features stacked U-structures for latent feature extraction. Such a design is based on the current state-of-the-art natural image segmentation model[40], allowing extensive multi-scale features extraction at relatively low computation and memory costs. Specifically, P-FEN is formed by five U-blocks that contain convolutional layers, batch-normalization layers, ReLU layers, and max-pooling layers (if necessary). The feature vectors derived from the global max-pooling layer are fed into a 1-layer perceptron for feature fusion, followed by a soft-max process to generalize normalized probability values. The final prediction probabilities of a disease label are then generated. The detailed design of the P-FEN is described in Supplementary Table S1.

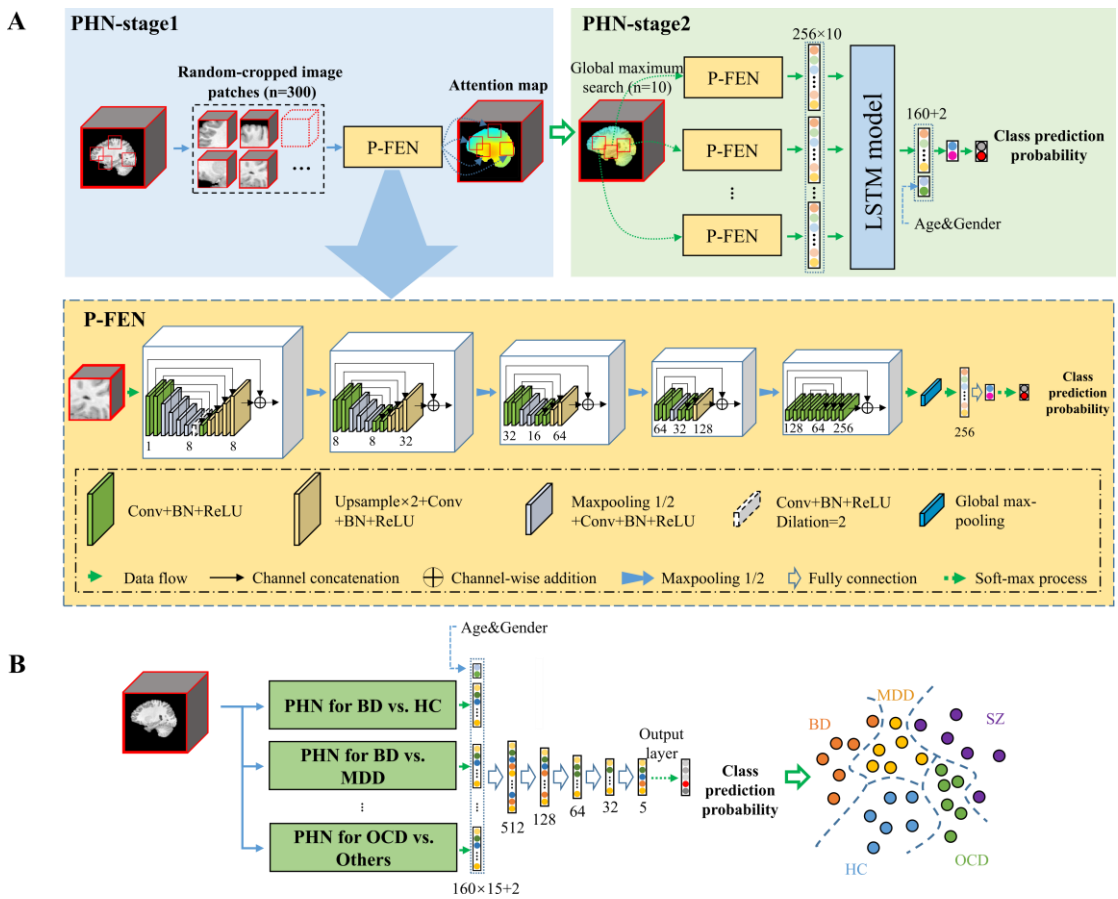


Figure 1. (A) The workflow of PHN. PHN-stage1 randomly samples image patches and input them to the P-FEN. The outputs of P-FEN form an attention map. PHN-stage2 focuses on the image patches with the top confidence and applies P-FEN and LSTM models to fuse the features of the top 10 spatially separated patches to make a final prediction. The detailed structure of the P-FEN, which is the core engine of the PHN, is also shown in (A). P-FEN accepts a 3-dimensional image block as input and obtains the disease probability prediction after several feature extraction filters such as convolutional and pooling layers. The numbers in each U-block are the number of output channels of the input layer, output layer, and hidden layer. (B) Model structure of the PHN multi-disease classifier.

PHN-stage1 randomly extracts 300 image patches with a size of 48×48×48 from the input image and feeds them into the trained P-FEN model to obtain disease probabilities. The centers of the patches are constrained within the brain. Based on the predicted probabilities of the 300 patches, an individual-specific confidence map (i.e., the attention map) is generated (Methods and Supplementary Figure S1). According to the attention map, PHN-stage2 selects ten image patches centered on voxels with peak disease prediction confidence. The algorithm only selects image patches that do not cover the center of each other to avoid information redundancy. The selected image patches are ranked by their prediction confidence and input to the P-FEN model. The feature vectors derived from the P-FENs (behind the global max-pooling layer), are

concatenated into a 256×10 matrix. The two dimensions of the matrix reflect local features in a given patch and the order of the spatially separated patches, respectively. This information is then processed by a long-short-term memory (LSTM) model[41], which considers the inter-dependence of the spatially separated patches as a sequence. The LSTM cell in the model has two hidden layers with a dropout rate of 0.5, and each layer has 16 neurons. LSTM generates a 160-dim feature vector that combines local features of multiple locations in the brain. Age and gender are further concatenated into the output of LSTM, and the final disease prediction probability is derived through a 1-layer perceptron. In order to unify the scale of age and gender with that of the features from LSTM, the value of age is divided by 100, and the gender is set to 0.4 for males and 0.6 for females.

PHN classifiers accurately recognize one diagnosis from the others

Based on PHN, we trained and tested a serial of image classifiers using T1-weighted 3D brain structural images from 2490 individuals, including 468 BD, 169 MDD, 1285 SZ, 268 OCD, and 300 HC. The data we used were from the Shanghai Mental Health Center Psychiatric Imaging Consortium (PIC), which was contributed by over 10 research groups with high-quality, carefully labeled neuroimaging data that acquired on the same scanner with minimal scan parameter differences (see Methods). Figure 2 presents the performance of PHN classifiers in five-fold cross-validation. The PHN classifiers achieved ⩾0.90 of the area under the ROC curve (AUC) in 7 of 10 pair-wise classification tasks (BD-HC, BD-MDD, BD-OCD, SZ-HC, SZ-MDD, SZ-OCD, and OCD-HC)(Figure 2A). The lowest AUC appeared in classifying BD and SZ (0.68). More useful classification tasks for clinical practice are recognizing a specific diagnosis type from the others. PHN classifiers achieved 0.90, 0.73, 0.91, 0.84, and 0.93, respectively, for recognizing HC, BD, MDD, SZ, and OCD individuals (Figure 2B). When comparing to other classification algorithms, the PHN classifiers showed higher AUC in all the 15 tasks (Figure 3, Supplementary Table S2).

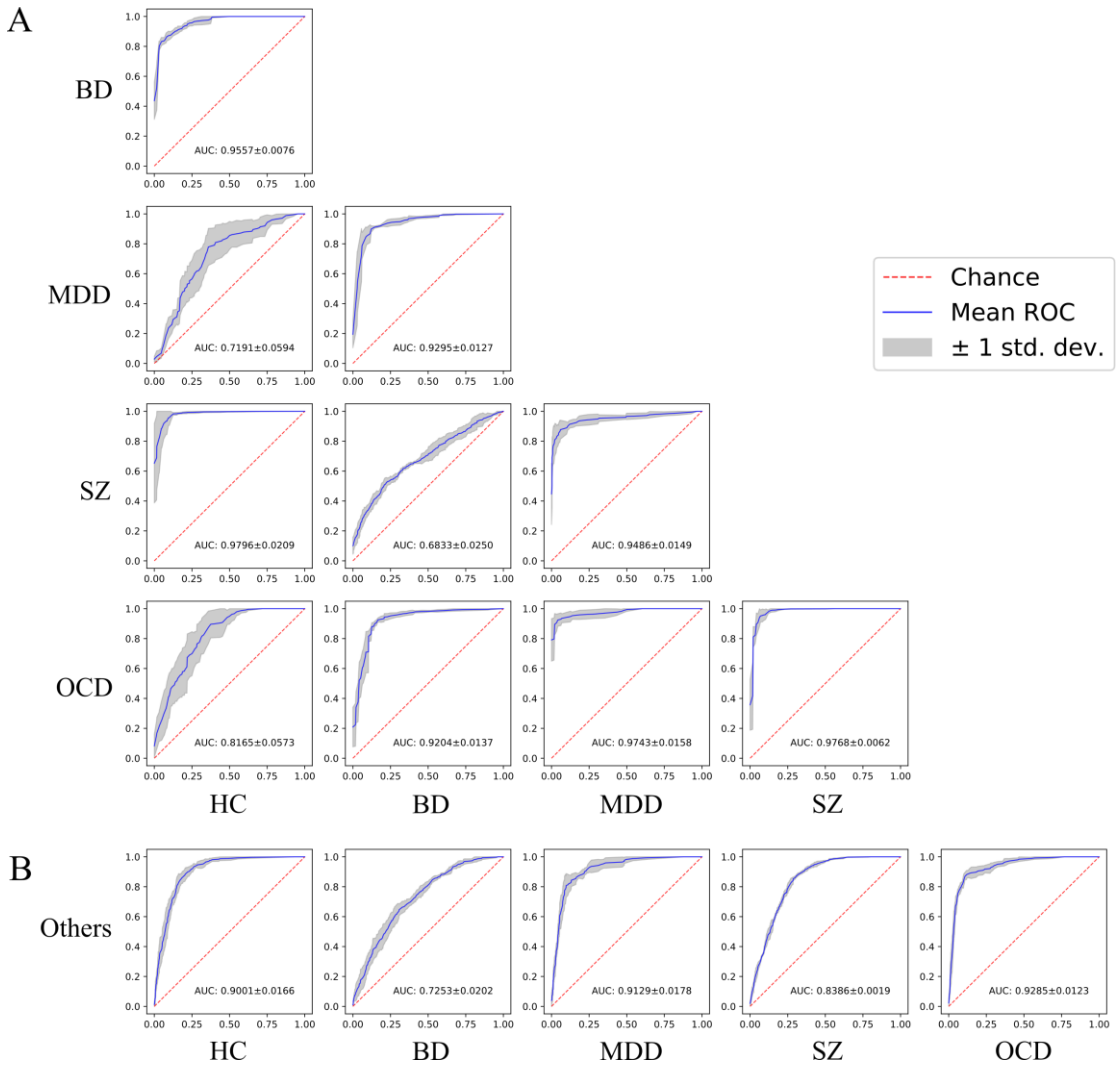


Figure 2. ROC curves of PHN for classifying HC, BD, MDD, SZ, and OCD. (A) ROC curves for classifying all possible pairs of diagnoses. The blue curves represent the averaged ROC, and the grey bands represent the standard deviation of the ROC curves among the five-fold cross-validations. (B) ROC curves of PHN for recognizing specific diagnoses from the others. AUC: area under the ROC curve, presented as mean ± standard deviation.

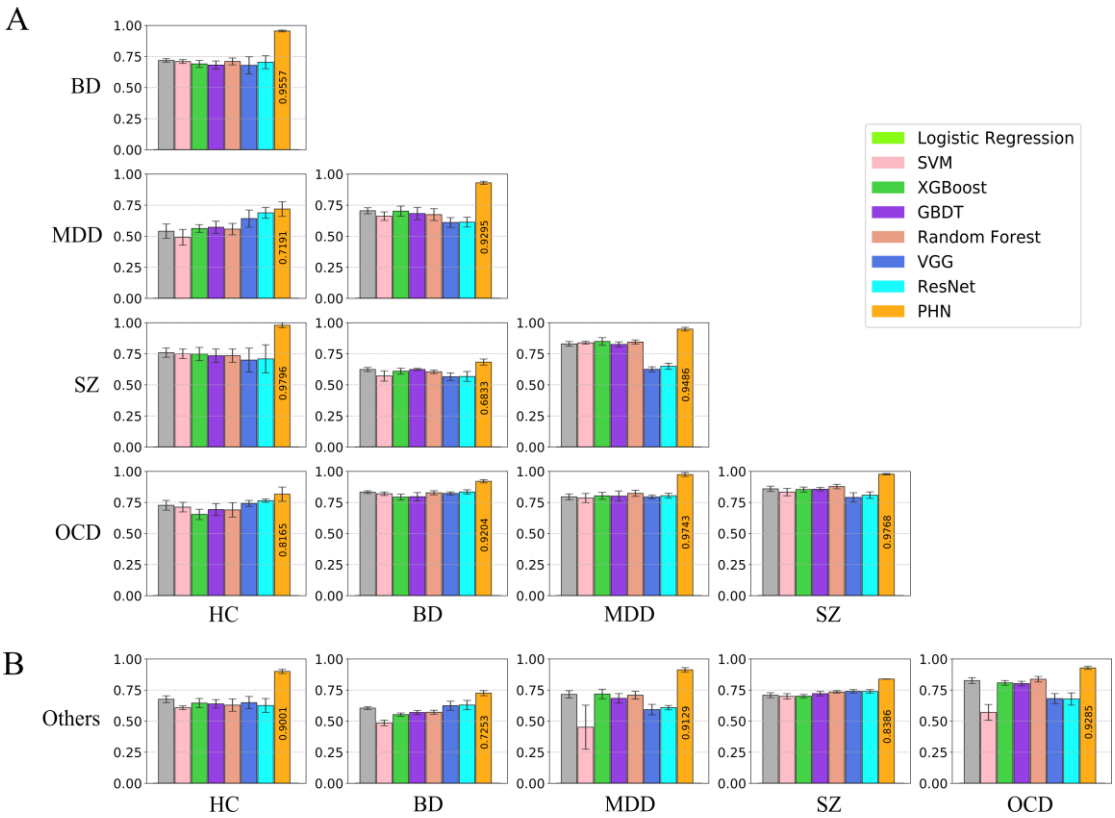


Figure 3. Comparisons of the area under the ROC curve (AUC) between PHN and the other algorithms. (A) All possible pairs of diagnoses; (B) Recognizing specific diagnoses from the others. SVM: support vector machine; XGBoost: extreme gradient boosting, GBDT: gradient boosting decision tree; VGG: 3D visual geometry group network; ResNet: 3D residual network; PHN: patch-based hierarchical network

The structure of PHN is essential for accurate prediction

We evaluated the effectiveness and robustness of PHN design via a series of ablation studies based on the HC vs. Others task. First, replacing the PHN-stage1 as a random patch selection module reduced the performance of the classifier, as reflected by lower AUC (Figure 4A), supporting the effectiveness of the pre-attention stage. Second, replacing the P-FEN module using other commonly used neural network models, such as VGG and ResNet, degraded the classification performance (Figure 4B), supporting the essential role of P-FEN. Third, replacing the LSTM model in PHN-stage2 with a multi-layer perceptron (PHN_linear in Supplementary Table S2) caused drop of AUC on 12 of all 15 classification tasks (Supplementary Table S2). These findings confirm the designated roles of the modules in PHN.

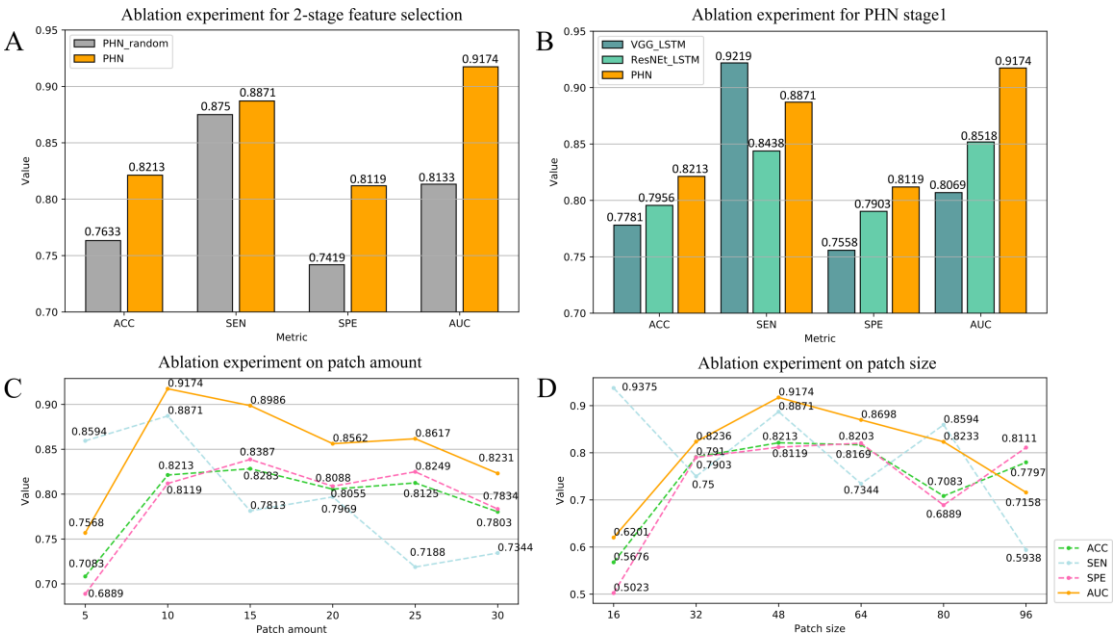


Figure 4. Results of ablation experiments based on the HC-others classification task. (A) The performance difference between randomly taking image blocks in PHN-stage2 (PHN_random) and taking the image blocks according to the confidence order; (B) The performance difference between replacing the P-FEN module with other deep learning network models (PHN_VGG and PHN_ResNet) and the original PHN model; (C) The relationship between performance and the number of input image blocks; (D) The relationship between performance and the size of image blocks. ACC: accuracy; SEN: sensitivity; SPE: specificity; AUC: area under the ROC curve.

We further examined the robustness of PHN to the size (default: 48×48×48) and amount of selected image patches in PHN-stage2 (default: top 10 image patches with the highest classification confidence). Figure 4C reflects that including too few image patches (e.g., top 5) could reduce the performance of PHN. However, when including a sufficient number of image patches, the AUC of PHN remained relatively stable. Figure 4D shows that a range of patch sizes, from 32 to 80, yielded a high AUC of PHN, suggesting the robustness of the algorithm against different patch size parameters.

Five-category classifications of mental diseases

In the rare cases that require five-alternative choices, PHN could combine the features extracted by the PHN binary classifiers (Figure 1B) to predict the disease label. Each image was input into the 10 PHN binary classifiers between different diagnosis labels and the 5 PHN one vs. others recognition models for feature extraction. The selected features as well as age and gender were concatenated and fed into a 5-layer perceptron to obtain the final 5-class classification results.

Figure 5A shows that the sensitivity metrics for identifying each of the five categories were higher than the random-guess probability (0.12, 0.19, 0.07, 0.52, and 0.11 for HC, BD, MDD, SZ, and OCD, respectively). The overall accuracy was also higher than 50%. The distribution of sensitivity and accuracy was stable among the

five-fold cross-validation. The confusion matrix (Figure 5B) indicates that the 5-categroty classifier correctly identified most individuals in each category (the highest proportion in each column is on the diagonal), and the error was due to difficulties in distinguishing specific category pairs, such as OCD-HC, BD-SZ, and MDD-HC. This error pattern suggests the potential to further enhance the performance by combining disease-specific classifiers with different advantages.

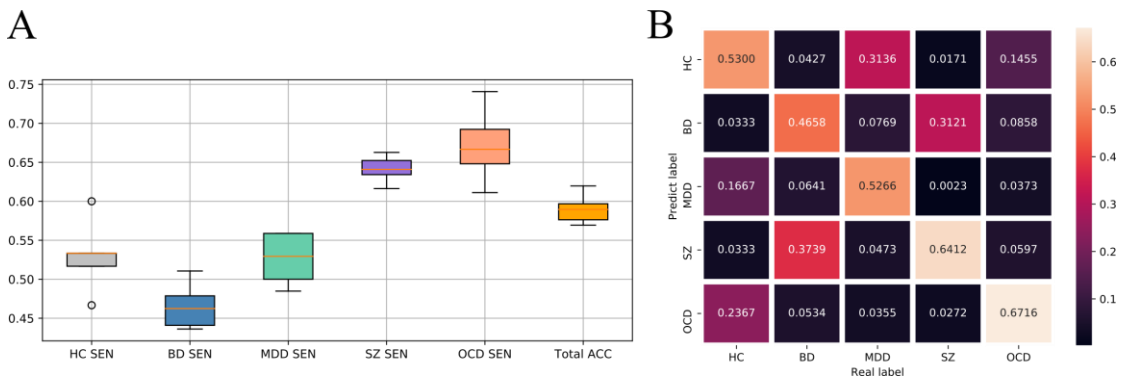


Figure 5. The performance of the 5-category classification model. (A) Sensitivity of each category and total accuracy. The box plot represents the variation in the five-fold cross-validation. (B) Confusion matrix of the classification results. The columns of the matrix represent the actual group labels. Each element in the matrix represents the proportion of correctly identified individuals to the total number of people in the group represented by the column. The proportions of each column add up to 1. ACC: accuracy; SEN: sensitivity.

Generalizable performance in independent validation

Table 1 shows the performance of PHN classifiers on four open data sets (LA5c[42], COBRE[43-45], HBN[46], and SRPBS[47]) (see Methods). In the LA5c, COBRE, and HBN datasets, the AUCs of BD-HC and SZ-HC classification tasks were above 0.8, and the AUC of MDD-HC and BD-SZ tasks was lower. In the SRPBS dataset, the AUC of the MDD-SZ classification task was 0.69, while the AUC of other tasks was above 0.78. These results support the generalizability of PHN and echo the relative strengths and weaknesses reflected in the cross-validation study.

Table 1. Performance of PHN on four open datasets.

Dataset	Task	ACC	SEN	SPE	AUC
LA5c (120 HC, 44 SZ, 48 BD)	BD vs HC	0.7815	0.5641	0.8875	0.8251
	SZ vs HC	0.7826	0.6	0.8625	0.8068
	BD vs SZ	0.527	0.6154	0.4286	0.5563
COBRE	SZ vs HC	0.8053	0.8364	0.7759	0.8335

(73 HC, 69 SZ)					
HBN (208 HC, 64 MDD, Children)	MDD vs HC	0.633	0.6471	0.6287	0.6467
SRPBS (926 HC, 38 BD, 237 MDD, 145 SZ)	BD vs HC	0.8091	0.6	0.8176	0.8217
	BD vs MDD	0.7955	0.6667	0.8158	0.7896
	BD vs SZ	0.7329	0.6333	0.7586	0.8037
	MDD vs HC	0.7882	0.7421	0.8	0.8154
	MDD vs SZ	0.6601	0.6158	0.7328	0.6936
	SZ vs HC	0.7325	0.7672	0.727	0.7825

Abbreviations: LA5c: UCLA Consortium for Neuropsychiatric Phenomics LA5c dataset; COBRE: Centers of Biomedical Research Excellence dataset; HBN: the Healthy Brain Network dataset; SRPBS: the Japanese Strategic Research Program for the Promotion of Brain Science multi-disorder MRI dataset (restricted); ACC: accuracy; SEN: sensitivity; SPE: specificity; AUC: area under the ROC curve.

We also compared PHN classifiers' performance with those published in previous studies[31-32,48-55] (Figure 6). PHN obtained the highest AUC in the SZ vs. HC, OCD vs. HC, and BD vs. HC task in classification performance.

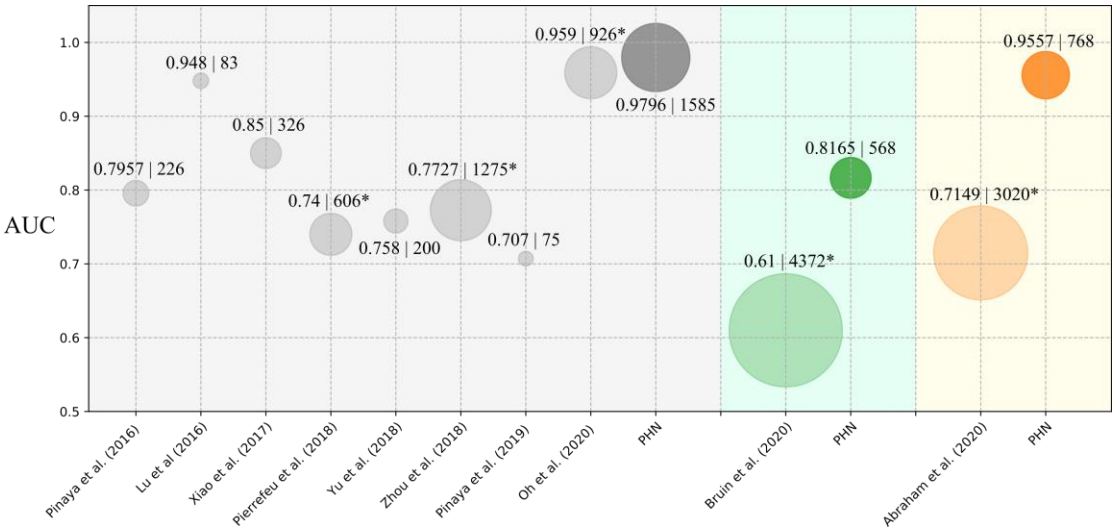
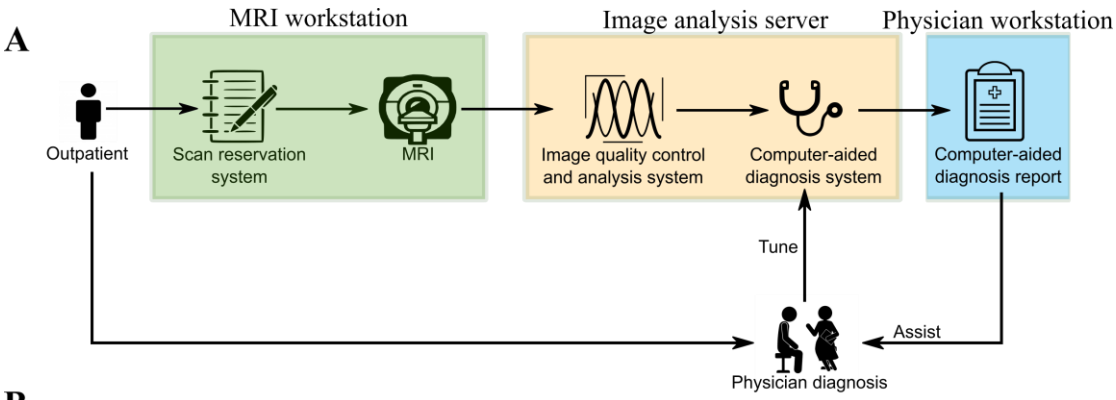


Figure 6. Comparison of the experiments in this paper with some existing work on AUC. Each circle indicates a participating work for comparison, blue indicates SZ vs. HC, orange indicates OCD vs. HC, and green indicates BD vs. HC. The circle's radius indicates the sample size of the study, and the data label next to the circle is (AUC value | sample size). The * superscript of sample size refers that the study included data from

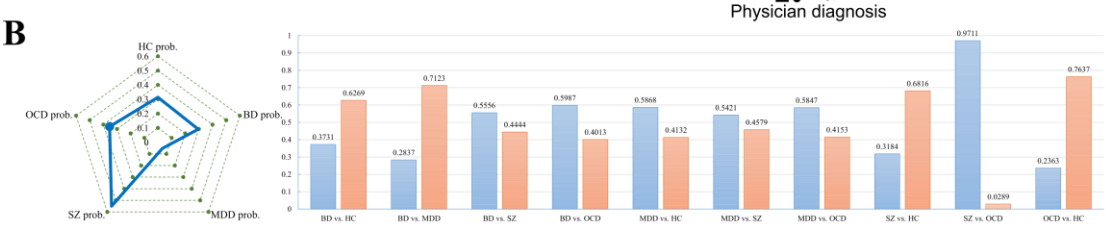
multiple scanning centers. The vertical axis indicates the AUC value obtained for each work.

Real world implementation and validation

Based on the PHN classifiers, we implemented a functioning application for neuroimaging-aided diagnosis of major mental disorders in Shanghai Mental Health Center. Figure 7A presents a scheme of the system, which automatically connects the physician workstation, MRI workstation, image analysis server. This system achieved a closed loop of psychiatrist's request, image acquisition, automatic analysis, result delivery, and psychiatrist's feedback. The PHN classifiers can be further tuned based on the feedback of psychiatrists' consensus diagnoses. Besides the output of PHN classifiers, this system also conveys the detected brain regions that deviate from the normal range in healthy population in terms of morphological indices (e.g., gray matter volume) and the extent to which these abnormalities are associated with disease diagnoses (odds ratio) (Figure 7B).



B



(a) (b)

Hemisphere	ROI	Volume (mm³)	Proportion of intracranial volume (%)	Higher/lower	General range of healthy people (%)	BD odds ratio*	MDD odds ratio	SZ odds ratio	OCD odds ratio
Left hemi.	Left banks of superior temporal sulcus	2597.00	0.168		0.128~0.235				
Left hemi.	Left caudal anterior cingulate cortex	1800.00	0.117		0.076~0.161				
Right hemi.	Right caudal middle frontal gyrus	4711.00	0.305	lower	0.326~0.523	2.037	1.346	1.764	1.393
Right hemi.	Right cuneus cortex	3082.00	0.200		0.162~0.268				
Right hemi.	Right entorhinal cortex	1667.00	0.108		0.093~0.181				

(c)

C

Inital diagnosis from clinic	Final diagnosis	BD vs. Others	MDD vs. Others	SZ vs. Others
Psychiatric disorders to be determined	Bipolar depression (mixed episodes)	0.5447	0.2014	0.4989
Psychiatric disorders to be determined	Bipolar disorder, currently a major depressive episode with psychotic symptoms	0.4861	0.2206	0.5173
Psychiatric disorders to be determined	Bipolar depression (mixed episodes)	0.5848	0.4958	0.5239
Psychiatric disorders to be determined	Bipolar disorder, currently a manic episode with psychotic symptoms	0.5212	0.063	0.5944
Excited and agitated state	Bipolar disorder	0.5942	0.0513	0.4733
Manic episodes	Bipolar disorder, currently a manic episode with psychotic symptoms	0.4985	0.0423	0.6046
Bipolar disorder	Bipolar disorder, currently a manic episode without psychotic symptoms	0.5202	0.0454	0.5431
Bipolar disorder	Bipolar disorder, currently a manic episode without psychotic symptoms	0.5324	0.2164	0.4619
Depressive state	Bipolar disorder, currently a major depressive episode without psychotic symptoms	0.4718	0.1917	0.4726
Bipolar disorder	Bipolar disorder, currently a manic episode with psychotic symptoms	0.6023	0.0931	0.5066
Bipolar disorder	Bipolar disorder, currently a manic episode with psychotic symptoms	0.5195	0.0557	0.5426
Excited and agitated state	Bipolar disorder, currently a manic episode without psychotic symptoms	0.6535	0.3435	0.4066
Bipolar disorder	Bipolar disorder	0.6505	0.0495	0.5101
Psychiatric disorders to be determined	Recurrent depressive disorder	0.4122	0.3297	0.4653
Depressive state	Recurrent depressive disorder, currently a major episode without psychotic symptoms	0.2649	0.5466	0.4567
Depressive state	Recurrent depressive disorder, currently a major episode without psychotic symptoms	0.4529	0.5906	0.4224
Hysterical psychosis	Major depressive episode with psychotic symptoms	0.4949	0.0592	0.5614
Depressive state	Major depressive episode with psychotic symptoms	0.4564	0.0949	0.5601
Depressive state	Recurrent depressive disorder, currently a major episode without psychotic symptoms	0.1317	0.6528	0.45
Organic mental disorders	Schizophrenia	0.5548	0.6738	0.4079
Schizophrenia	Schizophrenia	0.1456	0.7452	0.5353
Paranoid state	Schizophrenia	0.5054	0.034	0.5041
Psychiatric disorders to be determined	Schizophrenia	0.1016	0.3298	0.5566
Psychiatric disorders to be determined	Schizophrenia	0.271	0.4298	0.5274
Psychiatric disorders to be determined	Schizophrenia	0.3624	0.6242	0.4358
Psychiatric disorders to be determined	Schizophrenia	0.0857	0.5492	0.5079
Psychiatric disorders to be determined	Schizophrenia	0.0733	0.0726	0.5346
Paranoid state	Schizophrenia	0.4253	0.1045	0.5692
Schizophrenia	Schizophrenia	0.4798	0.0635	0.5553
Bipolar disorder	Schizophrenia	0.4756	0.0306	0.5931
Schizophrenia	Schizophrenia	0.3658	0.3032	0.5572
Schizophrenia	Undifferentiated schizophrenia	0.1618	0.0341	0.4809
Psychiatric disorders to be determined	Schizophrenia	0.4605	0.051	0.53
Bipolar disorder	Adolescent schizophrenia	0.5685	0.1277	0.5102
Schizophrenia	Paranoid schizophrenia	0.3465	0.1911	0.5157
Excited and agitated state	Paranoid schizophrenia	0.5409	0.0334	0.5881
Schizophrenia	Paranoid schizophrenia	0.4375	0.0309	0.5461
Schizophrenia	Schizophrenia	0.3881	0.0702	0.5928
Psychiatric disorders to be determined	Schizophrenia	0.3826	0.0999	0.5307
Acute and transient psychotic disorder	Paranoid schizophrenia	0.4853	0.1916	0.3679
Excited and agitated state	Schizophrenia	0.3694	0.0714	0.5551
Schizophrenia	Schizophrenia	0.4904	0.1892	0.5125
Excited and agitated state	Schizophrenia	0.4437	0.5659	0.4618
Schizophrenia	Schizophrenia	0.4725	0.0408	0.5695
Schizophrenia	Schizophrenia	0.2572	0.1198	0.5663
Emotional disorder	Schizophrenia	0.536	0.0301	0.4414
Schizophrenia	Schizophrenia	0.0842	0.1733	0.5774
Schizophrenia	Schizophrenia	0.3826	0.0371	0.5141
Schizophrenia	Undifferentiated schizophrenia	0.4077	0.0458	0.5734
Excited and agitated state	Undifferentiated schizophrenia	0.429	0.1439	0.4879
Schizophrenia	Undifferentiated schizophrenia	0.247	0.0855	0.5614
Schizophrenia	Schizophrenia	0.4997	0.2595	0.5721
Schizophrenia	Schizophrenia	0.4698	0.065	0.5678
Schizophrenia	Schizophrenia	0.1779	0.4813	0.4217
Schizophrenia	Schizophrenia	0.5958	0.1467	0.5455
Schizophrenia	Schizophrenia	0.4518	0.4865	0.5443
Excited and agitated state	Schizophrenia	0.4458	0.2007	0.5022
Hallucinatory delusional state	Schizophrenia	0.3837	0.0314	0.5996
Schizophrenia	Catatonic schizophrenia	0.6245	0.0473	0.4936
Schizophrenia	Schizophrenia	0.4876	0.1192	0.5971
Excited and agitated state	Schizophrenia	0.0542	0.519	0.5362
Bipolar disorder	Schizophrenia	0.4244	0.0571	0.546

Figure 7. The framework and application of the neuroimaging-aided diagnosis system

of major mental disorders in Shanghai Mental Health Center. A: The overall architecture of the system. B: Examples of analysis results that can be acquired by the neuroimaging-aided diagnosis system. (a) Recognition probability distribution of one diagnosis from the others. (b) Transdiagnostic disease prediction probability; (c) Volumetric status, as well as the dominance ratio for various psychiatric disorders of the different brain regions of a certain patient. C: PHN's performance in a real-world application with 62 patients. The first column presents the initial (transitional) diagnosis from the clinic. The second column presents the consensus diagnosis from a team of psychiatrists based on a series of examinations and observations in the ward. Orange, green, and blue code the diagnostic categories of BD, MDD, and SZ, respectively. The last three columns show the confidence values obtained by the three PHN classifiers. The confidence value> 0.5 is colored, and the transparency of the color encodes the confidence value. *DISEASE TYPE* is the exact diagnosis given by clinicians according to DSM-5 criteria.
* Odds ratio characterizes the degree of correlation between the indicator and the current disease. odds ratio>1: the indicator is positively correlated with the occurrence of the current disease; odds ratio<1: the indicator is negatively correlated with the occurrence of the current disease; odds ratio=1: the indicator is not correlated with the occurrence of the current disease.

Based on this clinical application, we validated the performance of PHN in real-world clinical practice. Based on T1-weighted 3D MR brain images of 323 patients (127 MDD, 83 BD, and 113 SZ) collected in the outpatient clinic, the AUC for classification of PHN was 0.6215 for BD-other task, 0.7654 for MDD-other task, and 0.6704 for SZ-other task (Supplementary Figure S2). In addition, we obtained consensus diagnoses for 62 of 323 patients (6 MDD, 13 BD, and 43 SZ), which were made by a team of psychiatrists based on a comprehensive examination and observations on the ward. Based on the consensus diagnosis and decision cutoff = 0.5, the classification accuracy of PHN was 0.8387 for BD-other tasks, 0.8548 for MDD-other tasks, and 0.6935 for SZ-other tasks. As shown in Figure 7C, a large portion of the errors made by the PHN classifier were accompanied by co-morbidities in the consensus diagnoses, such as "bipolar disorder with psychotic symptoms". The output of the PHN classifier echoes this co-morbidity phenomenon, suggesting that the classifier correctly reflects the complexity of psychiatric diagnoses in the real world.

Discussion

The complex symptoms of mental disorders and the lack of objective diagnostic tools make psychiatric diagnosis difficult. Here, we demonstrate for the first time that the combination of structural brain images with deep learning network models can help solve the problem of diagnosing multiple major psychiatric disorders in the real world. We propose a deep learning network, PHN, and apply it to a large-scale cross-disease

brain imaging dataset to train a series of classifiers that can identify any of the classes of HC, MDD, BD, SZ, and OCD. These classifiers showed good performance on cross-validation, independent public data, and real-world clinical data. This study achieves the ability to accurately identify one of multiple diagnoses and allows multiple diagnoses to coexist, thus addressing a real-world need in psychiatric practice. Based on these efforts, we have implemented a closed-loop system for neuroimaging-assisted diagnosis of major psychiatric disorders and have initially validated its effectiveness in real-world applications. Thus, this research contributes to the translation of brain research into an objective diagnostic aid for psychiatric disorders.

Identify specific diagnoses among multiple mental disorders and allow for comorbidities

A novel feature of the PHN adapted to the needs of clinical applications is its ability to accurately identify one type of diagnosis from among multiple types of psychiatric disorders (Figure 2B). This feature is better suited to address the practical clinical needs of psychiatrists who need to weigh multiple diagnoses than the results of most previous studies that focused on two groups. Meanwhile, the PHN classifiers enable parallel inference of different diagnostic categories, thus allowing more than one diagnostic inference for the same individual. This feature also makes the output of the PHN classifier more consistent with the highly co-morbid clinical characteristics of mental disorders[4,6]. Besides, PHN classifiers also perform well on most 2-category classification tasks, so psychiatrists can only refer to the PHN classifiers' recommendations for competing diagnostic types.

PHN's design and the cross-disease training dataset help improve classifier performance

Most of the machine learning studies on brain structure imaging of SZ, MDD, BD, and OCD combined conventional classifiers (such as support vector machine, SVM) and hand-crafted image features (e.g., voxel-based morphometrics, regional gray matter volume, and image features based on spatial deformation, tensor analysis, and shape analysis) [29-31,33-38]. Research in other fields[56-58] has shown that deep learning is more accurate than traditional neuroimaging classification algorithms. However, the few current studies applying deep learning methods have used hand-crafted image features[39] or skipped active feature selection -- treating the whole-brain image data equally[32]. Hand-crafted features may overwhelm fine-grained structural features of individuals due to spatial manipulation and averaging over regions. Accordingly, some studies have supported the advantages of voxel-level analysis over region-based analysis[59-60]. Yet, given the limited sample size in the psychiatric imaging field, equally treating the voxels of the whole brain may cause information redundancy and reduce training effectiveness. Therefore, deep learning methods with attention mechanisms are expected to help improve the performance of classifiers.

PHN is tailored to the specific needs of neuroimaging classification for mental disorders, resulting in PHN classifiers outperforming several other traditional machine learning and deep learning methods for all classification tasks (Figure 3). First, voxel-level features may represent physiological information and play essential roles in image feature-based classifiers[59-60], but aligning images to a standard space or averaging

signals within regions may disturb voxel-level features. PHN evaluates and selects local features (spatial patches) by training a deep learning model (P-FEN), thus avoiding the voxel-level interpolation and averaging process. Figure 4B provides empirical evidence for the effectiveness of P-FEN, and Figure 4C-D support that the performance of PHN is stable for the number and size of spatial patches, unless they are too small. Second, with limited sample size, equally applying whole-brain information to train classifiers may lead to redundancy and reduce the training efficiency[62-63]. The attention stage of PHN helps to reduce the information redundancy in classifier training. This advantage is supported by comparing two deep learning algorithms (VGG[64] and ResNet[65]) that process the whole brain image uniformly (Figure 3A-B). Figure 4A also shows the essential role of pre-selection and ranking of spatial patches in PHN-stage1. Third, the brain works as an integrated system, and local brain abnormalities are interrelated[98-101]. Thus, treating spatially distributed abnormalities as independent predictors may lose essential information. PHN's new application of the LSTM module solves this issue by treating the interdependence of spatially separated patches as a sequence. The results in Supplementary Table S2 show that the LSTM module in PHN-stage 2 accommodates an overall better performance than traditional fully connected neural networks.

In addition, training on a carefully labeled large cross-disease dataset obtained on the same scanner helps to improve the effectiveness of the PHN. The dataset we used is a large-scale dataset covering a wide range of mental disorders and contains data from carefully labeled and quality-controlled studies. The data were enrolled by 10 psychiatric study groups according to strict criteria, using essentially the same scanning process and parameters, undergoing fine quality control, and processed and managed using the same hardware and software platform[66]. Data from the same location has advantages in terms of image quality and consistency in socio-cultural context and understanding of diagnostic criteria, which helps to reduce noise in identifying specific diseases. Applying the PHN to data from multiple diagnoses also makes it easier for the PHN to identify features of specific diagnostic categories. The advantages of classifier design and training dataset facilitate the performance of the PHN classifiers, not only in cross-validation and independent validations, but also beyond previous studies[31-32,48-55] (Figure 6).

The performance of PHN classifiers is generalizable and effective in real world applications

The cross-site generalization of the PHN classifier is supported by the validation using four independent datasets (Table 2). This result shows that with some fine-tuning, the PHN classifier can be applied to neuroimaging data collected using different scanners. This feature is important because site effects have a significant impact on the systematic bias of image segmentation, thus hurting the cross-site generalizability of neuroimaging markers based on image segmentation[67]. Although researchers have been working to harmonize site effects in image segmentation, there is no consensus on an effective solution[68-72]. PHN classifiers do not require an explicit image segmentation step, but rather embed possible site effects in local features through deep learning. It has been suggested that machine learning methods may be effective in tuning site differences, and our empirical evidence supports the cross-site generality of PHN

classifiers. Since the sample size and diagnosis type of independent open datasets limited the amount of data used to tune the PHN, we expect that the cross-site performance of the PHN can be further improved with larger samples.

Importantly, the real-world validation shed light on the application of PHN to real-world clinical problems that are much more complex than those in case-control group comparison studies. Subject to many factors, the initial diagnosis in clinical practice is often symptom suggestive and different from the final consensus diagnosis. Furthermore, it is common for symptoms to overlap (e.g., a patient diagnosed with BD or MDD may also have concurrent psychotic symptoms, Figure 7C). These facts may increase the inconsistency between diagnoses made by different psychiatrists. PHN provides an objective measure and evidence-based reference for diagnosis. Although classification performance is lower in real-world data than in rigorously cleaned scientific data (especially when the initial diagnosis is used as the diagnostic label), in most cases, PHN results are consistent with the consensus diagnosis. For those patients with overlapping symptoms, the discriminant probabilities for each disease category from the PHN reflect co-morbidities accordingly. Thus, going beyond case-control studies based on clean diagnostic labels, a novel contribution of this study is to achieve the translation from neuroimaging biomarker research to clinical practice with acceptable accuracy.

In addition, the consistency of PHN with the initial diagnosis was verified on 323 real-world patients. As seen in Supplementary Figure S2, while the AUC greater than 0.5 was achieved in the identification tasks of BD, MDD, and SZ, the diagnostic performance of PHN was unsatisfactory when the initial diagnosis was considered as the gold standard. An important reason for these results is the possible confusion between diseases in the initial diagnosis. Another piece of evidence supporting this inference is that the ROC curve reveals that most patients with an initial diagnosis labeled MDD tend to be predicted by the PHN to be in the "other category": in practice, the initial diagnosis tends to be conservative (a patient will not be classified as having a serious mental illness such as BD or SZ at the outpatient clinic unless the specific symptoms are extremely obvious), so "MDD patients" used to test for PHN were actually likely to be in another mental illness category.

A closed-loop system of neuroimage-assistant diagnosis system for major mental disorders

With PHN as the core engine, we have implemented a closed-loop neuroimaging-assisted diagnostic system. We have integrated it in a clinical workflow to validate and improve the performance and stability of the system in a real-world setting (Figure 7). To our best knowledge, this is the first cross-disease neuroimaging-assisted diagnostic system translated into a clinical service test. This system not only provides objective diagnostic reference for psychiatrists, but also enables continuous learning from physician feedback and patient progression records to improve accuracy and upgrade predictive capabilities. Based on the accumulation of large amounts of data, the system also suggests individual reports of abnormalities at the brain region level and the strength of their association with disease diagnosis (Figure7A-B), thus providing psychiatrists with more intuitive and interpretable objective cues. We expect this

neuroimaging-assisted diagnostic system to provide a truly usable and objective artificial intelligence reference for psychiatric clinical services as it is continuously improved and validated.

Future directions

Our primary considerations for using high-resolution T1w brain structure images as a classification feature: 1) this imaging modality shows high reliability[73]; 2) mental disorders show abnormalities in this imaging modality[31-32,48-55]; and 3) imaging techniques are already available to help reduce their scan time and enable assessment of head movement levels during the scan, thus increasing the likelihood of successful acquisition of high-quality images. Future work could involve abnormalities in brain function as additional diagnostic features besides other brain structural modalities. Some functional brain imaging paradigms, such as naturalistic imaging[74], have been shown to have high reliability and are relatively easy to implement, which may help to evoke abnormalities in brain working patterns.

This work uses current psychiatric diagnostic systems as the "gold standard" for implementing an artificial intelligence system that could assist current clinical practice. However, there are also nosologies and theories in psychiatry that differ from clinical diagnostic systems, such as HiTOP[75] and RDoC[76], to address the high heterogeneity and co-morbidity in the current diagnostic system. Future research could explore neuroimaging systems that detect different symptom dimensions and infer the benefits of other treatment options, thus making neuroimaging-assisted diagnostic systems more functional.

In addition, the classification performance between some diagnostic categories in this work needs further improvement, such as the MDD vs. HC and BD vs. SZ tasks. These phenomena are consistent with previous findings[77-78], and thus they may suggest pathological ambiguity. Future studies could focus on improvements in terms of underlying theory and technology.

In conclusion, this study has developed and verified a neuroimaging model fitting the clinical needs of psychiatric diagnosis. Based on this model, we build a closed-loop neuroimaging-assistant diagnostic system embedded in the clinical process. The effectiveness and applicability of this system has been verified in the real world application.

Methods

SMHC dataset

T1-weighted high-resolution brain structural images of healthy controls (HC) and patients with BD, MDD, SZ, OCD diagnoses were collected in Shanghai Mental Health Center (SMHC) in China, using a Siemens 3.0 Tesla Verio scanner with highly similar acquisition parameters (Supplementary Table S3). The participants were initially recruited by ten research groups of SMHC from local clinics/communities for various research projects from April 2015 to May 2020. All participants (or legal guardians) signed written informed consent during the recruitment. The IRB committee of

Shanghai Mental Health Center approved this study and permitted the use of measures extracted from the coded data (IRB2020-15). The following inclusion criteria for the participants were applied: (1) 18-59 years old; (2) A diagnosis of SZ, OCD, BD, or MDD according to Diagnostic and statistical manual of mental disorders (DSM-5) [79], DSM-IV, and the International Classification of Diseases (ICD)-10[80]. A diagnosis of BD respectively based on the DSM-5 (n =271), DSM-IV (n = 61) , and the ICD-10 (n = 136); The patients were diagnosed with MDD using the DSM- IV (n = 169); The patients were diagnosed with SZ using the DSM-5 (n = 39), DSM-IV (n = 95), and the ICD-10 (n = 1151); The patients were diagnosed with OCD using the DSM-5 (n = 205), DSM-IV (n =63); (3) Completed a high-resolution T1-3D brain scan that passed the quality check.

The exclusion criteria were: (1) any current or lifetime Axis I psychiatric disorder comorbidity screened by Mini-International Neuropsychiatric Interview (MINI)[81]; (2) a history of substance abuse and serious physical disease. The healthy controls were recruited from local communities. They underwent a MINI interview to exclude any mental disorder and had no history of antipsychotic medication and serious physical diseases.

Senior radiologists checked all images included in this study to exclude obvious brain structural abnormalities and imaging artifacts. The quality of the images was further quantified using CAT12.6 toolbox[82] based on the resolution, intensity bias, and noise level. Images with a quality grade lower than 'B-' were excluded from further analyses. After quality control, a total of 2490 data of participants, including 300 HC subjects, 468 BD patients, 169 MDD patients, 1285 SZ patients, and 268 OCD patients, were included in the subsequent analyses (Table 2).

Table 1 The demographic characteristics of 2490 subjects.

Characteristics	HC	BD	MDD	SZ	OCD
Gender (M/F)	153/147	196/272	68/101	654/631	162/106
Age (Mean±Std)	35.91±14.47	32.98±14.83	31.02±14.46	39.78±15.61	29.22±9.40

M=male; F=female; Std=standard deviation.

The images were resampled into a resolution of 256×256×256 and corrected for intensity inhomogeneity using the N3 algorithm[83]. Then we performed skull-stripping using watershed algorithm[84] and a process of manually checking to ensure that the skull was completely removed. All the above processing steps were performed using the same hardware and software platform[66].

Open research datasets

We used four open datasets to examine the generalizability of PHN's performance. The datasets included: 1) the UCLA Consortium for Neuropsychiatric Phenomics LA5c (LA5c) dataset[42], which contains 120 HC subjects, 44 SZ subjects, and 48 BD subjects;

2) Centers of Biomedical Research Excellence (COBRE) dataset[43-45], which contains 73 HC and 69 SZ subjects, 3) the Healthy Brain Network (HBN) dataset (release 1-9)[46], which contains 64 MDD and 208 HC children and teenagers, and 4) the Japanese Strategic Research Program for the Promotion of Brain Science (SRPBS) multi-disorder MRI dataset (restricted)[47], which contains 926 HC subjects, 38 BD subjects, 237 MDD subjects, and 145 SZ subjects. Details of these datasets are presented in Supplementary Table S4, and the quality control process and preprocessing steps are exactly the same as the process described above. Due to the lack of an open neuroimaging dataset for OCD, the OCD-related performance was not examined. The quality control process and preprocessing steps were fully consistent with the process described earlier. To reduce the influence of scanner differences on the results, we randomly used 20% of each public dataset to pre-train (fine-tune) the parameters of PHN, and the rest of the data were used as test data.

Real-world validation data

The greatest value of machine learning models lies in the integration of high-throughput information for practical services, but most neuroimaging-based classification models of mental illness, including PHN, were trained and verified based on data collected in accordance with scientific research standards. Since most research standards exclude comorbidities, ambiguous diagnoses, etc., it is not sufficiently representative of the real-world clinical situation.

We verified the performance of PHN in real-world clinical practice. We collected T1-weighted 3D MR brain images of 323 patients (83 BD, 127 MDD and 113 SZ, Supplementary Table S5), and their (interim) diagnosis obtained in the outpatient clinic. In addition, 62 patients (6 MDD, 13 BD, and 43 SZ) among 323 patients have consensus diagnosis (according to DSM-5 criteria) made by a team of psychiatrists based on comprehensive examinations and observations in the ward. The quality control process and preprocessing steps are the same as the process described above. PHN predicts the diagnosis based on brain images and evaluates real-world performance based on the interim and final diagnosis in the ward.

Patch-based hierarchical network (PHN)

As demonstrated in Figure 1A, we developed a deep learning framework, patch-based hierarchical network (PHN), to recognize the images' diagnosis labels. PHN contains two stages: PHN-stage1 detects diagnosis-sensitive brain locations, and PHN-stage2 performs local feature extraction and classification. The rationale behind this design is that pre-locating possible disease-related brain regions (PHN-stage1) may help reduce redundant information and improve classification performance (PHN-stage2).

PHN-stage1 randomly extracts 300 image patches with a size of $48 \times 48 \times 48$ from the input image and feeds them into the trained P-FEN model to obtain disease probabilities. The centers of the patches are constrained within the brain. Based on the predicted probabilities of the 300 patches, an individual-specific confidence map (i.e., the attention map) is generated. Specifically, the diagnosis confidence with respect to the i-th voxel in the brain image has the following value $\mathcal{C}_i$:

$$\mathcal{C}_i = \frac{1}{K_i}\left[\sum_{k=1}^{K_i} \mathcal{M}_k(\|0.5 - \mathcal{P}(x_k)\| + 0.5)\right]_i$$

where K_i denotes the frequency of the i-th voxel in the 300 patches, $\mathcal{M}_k(t): [0.5, 1] \rightarrow \mathbb{R}^{256\times256\times256}$ refers to the operator that adds the value of $t \in [0.5, 1]$ to the patch with an appropriate location in a zero-padded image in $\mathbb{R}^{256\times256\times256}$, and $\mathcal{P}(x_k) \in [0, 1]$ denotes the disease prediction probability for the k-th image patch. Here, $\|0.5 - \mathcal{P}(x_k)\|$ calculates the distance between 0.5 and $\mathcal{P}(x_k)$, as the prediction probability of 0.5 refers to the random guess while the probabilities close to 0 and 1 indicate the prediction with high confidence; 0.5 was further added to ensure the calculated confidence value ranges is between 0.5 to 1, which 0.5 indicates the absence of any categorical tendency, while 1 indicates a strong conviction of categorical tendency. Supplementary Figure S1 presents examples of such voxel-wise attention maps.

Model training and cross-validation

The PHN model was implemented using Python 3.7 based on PyTorch 1.6.0, and the computer we used contains a single GPU (i.e., NVIDIA Quadro P5000 16 GB), and the operating system is CentOS 7.9. Classifiers that distinguish every two groups among HC, BD, MDD, SZ, and OCD groups, as well as one group from the others, were trained separately. The classifiers were trained and tested using a 5-fold cross-validation scheme. The samples in each diagnosis group were divided into 80% of the training set and 20% of the testing set in each fold. Within the training set, 2/3 of the samples (keeping the ratio of the two groups) were used to train the P-FEN model for 300 epochs, and the other 1/3 samples were used to evaluate the performance and select the best model. The trained P-FEN was then deployed to PHN-stage2, and the same procedure and data as above were used to train PHN-stage2. For all training steps, the adaptive moment estimation algorithm (ADAM)[85] was used as the optimizer with a learning rate of 1×10^{-4}, and the batch size was set to 28.

Performance metrics

We computed sensitivity (SEN), specificity (SPE), the area under the receiver operating characteristic curve (AUC), F1-score (F1), and Matthews correlation coefficient (MCC) in each test set in the cross-validation. The F1-score and MCC are defined as:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{TP}}{(2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN})}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FN})}}$$

where TP and TN denote true positive and true negative values, and FP and FN refer to false positive and false negative values, respectively. The F1 and MCC are both measures for the classification performance of the 2-class classification task. F1 considers both precision and recall, and MCC serves as a balanced measure of classes

of different sample sizes. The mean and standard deviation values of these metrics were calculated among the five folds of cross-validation.

Comparisons with other methods

We compared the performance of PHN with five widely-used machine learning methods, including 1) logistic regression, with L2 penalty, C=1, and 1e-4 of tolerance for stopping criteria; 2) support vector machine (SVM)[86], with linear kernel, C=1, and 1e-3 of stopping tolerance; 3) random forest classifier[87], with 100 trees, Gini impurity as split criteria, 2 for the minimum number of samples required to split an internal node, and 1 for the minimum number of samples required to be at a leaf node; 4) extreme gradient boosting (XGBoost)[88], with maximum tree depth=2, learning rate=0.1, estimator number=100; 5) gradient boosting decision tree (GBDT)[89], with deviance as loss, 0.1 as learning rate, estimator number=100, Friedman-MSE as a criterion.

To derive imaging features for these methods, we used FreeSurfer (version 6.0.0)[90] to compute volume, cortical thickness (for cortical structures only), and cortical surface area (for cortical structures only) measures for brain regions (87 regions) defined in the Desikan-Killiany atlas[91]+subcortical atlas[92]. In addition, we extracted 3D image histological features for the 87 brain regions by using Radiomics (version 3.0.1)[93]. To avoid overfitting, the two sets of features were separately reduced to eight dimensions by principal component analyses and combined into a 16-dimensional feature set. The machine learning models above were then trained using this feature set.

We further compared PHN with two other deep learning models: a 3D visual geometry group network (VGG)[64], which has been widely used in brain disease diagnosis[94-97], and a 3D residual network (ResNet)[65], which has been proved to be the most effective convolutional neural network in computer vision tasks. We slightly modified the structure of VGG and ResNet (see Supplementary Table S6-7) to make them similar to the networks used in existing studies[23-25] and to reduce the computational cost.

Ablation studies on the framework structure

To examine the effectiveness of PHN, we performed a series of ablation studies on the network structure. First, to verify the effectiveness of the two-stage framework (i.e., the first stage for brain region localization and the second stage for feature extraction and classification), we trained another 2-stage PHN model for the HC vs. Others task. In this model, ten image patches in PHN-stage2 were randomly selected (rather than selected based on prediction confidence) for further feature extraction. Second, to examine the effectiveness of the P-FEN structure, we replaced the P-FEN with the aforementioned VGG and ResNet in PHN. Third, since we specified the size of the image patches ($48\times48\times48$) for P-FEN and the number of image patches (top 10) in PHN-stage2, we conducted a series of experiments to examine the robustness of PHN to the size and amount of the image patches. Fourth, in order to verify the effectiveness of the LSTM model in PHN-stage2, we replaced the LSTM model with a multi-layer perceptron. Finally, we examined the effect of including age and sex information by removing the age and sex input in the final perceptron in PHN-stage2.

Multi-class classification

We further evaluated the capability of PHN to solve the five-class problem by combining the features extracted by different PHN classifiers (Figure 1B). Each image was input into the ten aforementioned PHN binary classifiers between different diagnosis labels and the 5 PHN recognition model for each disease label (i.e., the current label vs. others). Each of the 15 PHN models extracted a 160-dimensional feature set, and these sets, as well as age and gender information, were concatenated. Then, the concatenated features (dimension is 160×15+2=2402) were fed into a 5-layer perceptron to obtain the final 5-class classification results.

References

1. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness[J]. The Lancet Psychiatry, 2016, 3(2): 171-178.

2. Huang Y, Wang YU, Wang H, et al. Prevalence of mental disorders in China: a cross-sectional epidemiological study[J]. The Lancet Psychiatry, 2019, 6(3): 211-224.

3. Clark LA, Cuthbert B, Lewis-Fernández R, Narrow WE, Reed GM. Three Approaches to Understanding and Classifying Mental Disorder: ICD-11, DSM-5, and the National Institute of Mental Health's Research Domain Criteria (RDoC). Psychol Sci Public Interest. 2017 Sep;18(2):72-145. doi: 10.1177/1529100617727266. PMID: 29211974.

4. Sharma LP, Reddy YCJ. Obsessive-compulsive disorder comorbid with schizophrenia and bipolar disorder. Indian J Psychiatry. 2019 Jan;61(Suppl 1):S140-S148. doi: 10.4103/psychiatry.IndianJPsychiatry_527_18. PMID: 30745688; PMCID: PMC6343407.

5. Sharma V. Relationship of bipolar disorder with psychiatric comorbidity in the postpartum period—a scoping review[J]. Archives of women's mental health, 2018, 21(2): 141-147.

6. Upthegrove R, Marwaha S, Birchwood M. Depression and schizophrenia: cause, consequence, or trans-diagnostic issue?[J]. Schizophrenia bulletin, 2017, 43(2): 240-244.

7. Tomaiuolo F, Carlesimo G A, Di Paola M, et al. Gross morphology and morphometric sequelae in the hippocampus, fornix, and corpus callosum of patients with severe non-missile traumatic brain injury without macroscopically detectable lesions: a T1 weighted MRI study[J]. Journal of Neurology, Neurosurgery & Psychiatry, 2004, 75(9): 1314-1322.

8. Vijayakumar N, Allen N B, Youssef G, et al. Brain development during adolescence: A mixed-longitudinal investigation of cortical thickness, surface area, and volume[J]. Human brain mapping, 2016, 37(6): 2027-2038.

9. Westlye L T, Walhovd K B, Dale A M, et al. Differentiating maturational and aging-related changes of the cerebral cortex by use of thickness and signal intensity[J]. Neuroimage, 2010, 52(1): 172-185.

10. Gur R E , Calkins M E , Gur R C , et al. The Consortium on the Genetics of Schizophrenia: neurocognitive endophenotypes.[J]. Schizophrenia Bulletin, 2007, 33(1):49.

11. Smoller J W , Gardner-Schuster E . Genetics of bipolar disorder.[J]. Lancet, 2013, 4(6):22-29.

12. Kato T . Molecular genetics of bipolar disorder and depression[J]. Psychiatry and Clinical Neurosciences, 2007, 61(1):3-19.

13. Saveanu R V , Nemeroff C B . Etiology of Depression: Genetic and Environmental Factors[J]. The Psychiatric clinics of North America, 2012, 35(1):51-71.

14. Mcdonald C , Murray R M . Early and late environmental risk factors for schizophrenia[J]. Brain Research Reviews, 2000, 31(2):130-137.

15. Alloy L B , Abramson L Y , Urosevic S , et al. The psychosocial context of bipolar disorder: Environmental, cognitive, and developmental risk factors[J]. Clin Psychol Rev, 2005, 25(8):1043-1075.

16. Boyke J, Driemeyer J, Gaser C, et al. Training-induced brain structure changes in the elderly[J]. Journal of Neuroscience, 2008, 28(28): 7031-7035.

17. Kremen W S, Prom-Wormley E, Panizzon M S, et al. Genetic and environmental influences on the size of specific brain regions in midlife: the VETSA MRI study[J]. Neuroimage, 2010, 49(2): 1213-1223.

18. Jednoróg K, Altarelli I, Monzalvo K, et al. The influence of socioeconomic status on children’s brain structure[J]. 2012.

19. McIntosh A M, Job D E, Moorhead W J, et al. Genetic liability to schizophrenia or bipolar disorder and its relationship to brain structure[J]. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2006, 141(1): 76-83.

20. Peper J S, Brouwer R M, Boomsma D I, et al. Genetic influences on human brain structure: a review of brain imaging studies in twins[J]. Human brain mapping, 2007, 28(6): 464-473.

21. Basaia S, Agosta F, Wagner L, et al. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks[J]. NeuroImage: Clinical, 2019, 21: 101645.

22. Ahmed S, Choi K Y, Lee J J, et al. Ensembles of patch-based classifiers for diagnosis of Alzheimer diseases[J]. IEEE Access, 2019, 7: 73373-73383.

23. Liu M, Zhang J, Adeli E, et al. Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis[J]. IEEE Transactions on Biomedical Engineering, 2018, 66(5): 1195-1206.

24. Lian C, Liu M, Zhang J, et al. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 42(4): 880-893.

25. Qiu S, Joshi P S, Miller M I, et al. Development and validation of an interpretable deep learning framework for Alzheimer’s disease classification[J]. Brain, 2020, 143(6): 1920-1933.

26. Zhang J, Gao Y, Gao Y, et al. Detecting anatomical landmarks for fast Alzheimer's disease diagnosis[J]. IEEE transactions on medical imaging, 2016, 35(12): 2524-2533.

27. Liu M, Zhang D, Shen D, et al. Ensemble sparse classification of Alzheimer's disease[J]. NeuroImage, 2012, 60(2): 1106-1116.

28. Caspi A, Houts R M, Ambler A, et al. Longitudinal assessment of mental health disorders and comorbidities across 4 decades among participants in the Dunedin birth cohort study[J]. JAMA network open, 2020, 3(4): e203221-e203221.

29. Mwangi B, Ebmeier K P, Matthews K, et al. Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder[J]. Brain, 2012, 135(5): 1508-1521.

30. Madine M, Rekik I, Werghi N. Diagnosing Autism Using T1-W MRI With Multi-Kernel Learning and Hypergraph Neural Network[C]2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020: 438-442.

31. Lu X, Yang Y, Wu F, et al. Discriminative analysis of schizophrenia using support vector machine and recursive feature elimination on structural MRI images[J]. Medicine, 2016, 95(30).

32. Oh J, Oh B L, Lee K U, et al. Identifying schizophrenia using structural MRI with a deep learning algorithm[J]. Frontiers in psychiatry, 2020, 11: 16.

33. Schnack H G, Nieuwenhuis M, van Haren N E M, et al. Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects[J]. Neuroimage, 2014, 84: 299-306.

34. Rubin-Falcone H, Zanderigo F, Thapa-Chhetry B, et al. Pattern recognition of magnetic resonance imaging-based gray matter volume measurements classifies bipolar disorder and major depressive disorder[J]. Journal of affective disorders, 2018, 227: 498-505.

35. Hong S, Liu Y S, Cao B, et al. Identification of suicidality in adolescent major depressive disorder patients using sMRI: A machine learning approach[J]. Journal of affective disorders, 2021, 280: 72-76.

36. Tas C, Mogulkoc H, Eryilmaz G, et al. Discriminating schizophrenia and schizo-obsessive disorder: a structural MRI study combining VBM and machine learning methods[J]. Neural Computing and Applications, 2018, 29(2): 377-387.

37. Jie N F, Zhu M H, Ma X Y, et al. Discriminating bipolar disorder from major depression based on SVM-FoBa: efficient feature selection with multimodal brain imaging data[J]. IEEE transactions on autonomous mental development, 2015, 7(4): 320-331.

38. Fung G, Deng Y, Zhao Q, et al. Distinguishing bipolar and major depressive disorders by brain structural morphometry: a pilot study[J]. BMC psychiatry, 2015, 15(1): 1-12.

39. Korda A I, Ruef A, Neufang S, et al. Identification of voxel-based texture abnormalities as new biomarkers for schizophrenia and major depressive patients using layer-wise relevance propagation on deep learning decisions[J]. Psychiatry Research: Neuroimaging, 2021, 313: 111303.

40. Qin X, Zhang Z, Huang C, et al. U2-Net: Going deeper with nested U-structure for salient object detection[J]. Pattern Recognition, 2020, 106: 107404.

41. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

42. Poldrack R A, Congdon E, Triplett W, et al. A phenome-wide examination of neural and cognitive function[J]. Scientific data, 2016, 3(1): 1-12.

43. Calhoun V D, Sui J, Kiehl K, et al. Exploring the psychosis functional connectome: aberrant intrinsic networks in schizophrenia and bipolar disorder[J]. Frontiers in psychiatry, 2012, 2: 75.

44. Mayer A R, Ruhl D, Merideth F, et al. Functional imaging of the hemodynamic sensory gating response in schizophrenia[J]. Human brain mapping, 2013, 34(9): 2302-2312.

45. Hanlon F M, Houck J M, Pyeatt C J, et al. Bilateral hippocampal dysfunction in schizophrenia[J]. Neuroimage, 2011, 58(4): 1158-1168.

46. Alexander L M, Escalera J, Ai L, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders[J]. Scientific data, 2017, 4(1): 1-26.

47. Tanaka S C, Yamashita A, Yahata N, et al. A multi-site, multi-disorder resting-state magnetic resonance image database[J]. Scientific data, 2021, 8(1): 1-15.

48. Pinaya W H L, Gadelha A, Doyle O M, et al. Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia[J]. Scientific reports, 2016, 6(1): 1-9.

49. Pinaya W H L, Mechelli A, Sato J R. Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study[J]. Human brain mapping, 2019, 40(3): 944-954.

50. Xiao Y, Yan Z, Zhao Y, et al. Support vector machine-based classification of first episode drug-naïve schizophrenia patients and healthy controls using structural MRI[J]. Schizophrenia Research, 2019, 214: 11-17.

51. de Pierrefeu A, Löfstedt T, Laidi C, et al. Interpretable and stable prediction of schizophrenia on a large multisite dataset using machine learning with structured sparsity[C].2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI). IEEE, 2018: 1-4.

52. Yu W, Na Z, Fengxia Y, et al. Magnetic resonance imaging study of gray matter in schizophrenia based on XGBoost[J]. Journal of Integrative Neuroscience, 2018, 17(4): 331-336.

53. Zhou A, Cui Y, Jiang T. Multisite Schizophrenia Classification Based on Brainnetome Atlas by Deep Learning[C].2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS). IEEE, 2018: 451-455.

54. Bruin W B, Taylor L, Thomas R M, et al. Structural neuroimaging biomarkers for obsessive-compulsive disorder in the ENIGMA-OCD consortium: medication matters[J]. Translational psychiatry, 2020, 10(1): 1-12.

55. Nunes A, Schnack H G, Ching C R K, et al. Using structural MRI to identify bipolar disorders–13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group[J]. Molecular psychiatry, 2020, 25(9): 2130-2143.

56. Minaee S, Kalchbrenner N, Cambria E, et al. Deep learning--based text classification: a comprehensive review[J]. ACM Computing Surveys (CSUR), 2021, 54(3): 1-40.

57. Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 3642-3649.

58. Lundervold A S, Lundervold A. An overview of deep learning in medical imaging focusing on MRI[J]. Zeitschrift für Medizinische Physik, 2019, 29(2): 102-127.

59. Astrakas L G, Argyropoulou M I. Shifting from region of interest (ROI) to voxel-based analysis in human brain mapping[J]. Pediatric radiology, 2010, 40(12): 1857-1867.

60. Baecker L, Dafflon J, Da Costa P F, et al. Brain age prediction: A comparison between machine learning models using region-and voxel-based morphometric data[J]. Human brain mapping, 2021, 42(8): 2332-2346.

61. Rasmussen P M, Madsen K H, Lund T E, et al. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps[J]. NeuroImage, 2011, 55(3): 1120-1131.

62. Altman N, Krzywinski M. The curse (s) of dimensionality[J]. Nat Methods, 2018, 15(6): 399-400.

63. Tang Y, Chen D, Li X. Dimensionality reduction methods for brain imaging data analysis[J]. ACM Computing Surveys (CSUR), 2021, 54(4): 1-36.

64. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

65. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

66. Li Q, Jiang L, Qiao K, et al. INCloud: integrated neuroimaging cloud for data collection, management, analysis and clinical translations[J]. General Psychiatry, 2021, 34(6).

67. Glocker B, Robinson R, Castro D C, et al. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects[J]. arXiv preprint arXiv:1910.04597, 2019.

68. Zuo L, Dewey B E, Liu Y, et al. Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory[J]. NeuroImage, 2021, 243: 118569.

69. Pomponio R, Erus G, Habes M, et al. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan[J]. NeuroImage, 2020, 208: 116450.

70. Modanwal G, Vellal A, Buda M, et al. MRI image harmonization using cycle-consistent generative adversarial network[C]//Medical Imaging 2020: Computer-Aided Diagnosis. International Society for Optics and Photonics, 2020, 11314: 1131413.

71. Clarke W T, Mougin O, Driver I D, et al. Multi-site harmonization of 7 tesla MRI neuroimaging protocols[J]. NeuroImage, 2020, 206: 116335.

72. Guan H, Liu Y, Yang E, et al. Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification[J]. Medical Image Analysis, 2021, 71: 102076.

73. Hedges E P, Dimitrov M, Zahid U, et al. Reliability of structural MRI measurements: The effects of scan session, head tilt, inter-scan interval, acquisition sequence, FreeSurfer version and processing stream[J]. NeuroImage, 2022, 246: 118751.

74. Wang J, Ren Y, Hu X, et al. Test–retest reliability of functional connectivity networks during naturalistic fMRI paradigms[J]. Human brain mapping, 2017, 38(4): 2226-2241.

75. Kotov R, Krueger R F, Watson D, et al. The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies[J]. Journal of abnormal psychology, 2017, 126(4): 454.

76. Insel T, Cuthbert B, Garvey M, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders[J]. American Journal of psychiatry, 2010, 167(7): 748-751.

77. Fried E I, Kievit R A. The volumes of subcortical regions in depressed and healthy individuals are strikingly similar: a reinterpretation of the results by Schmaal et al[J]. Molecular psychiatry, 2016, 21(6): 724-725.

78. Madre M, Canales-Rodríguez E J, Fuentes-Claramonte P, et al. Structural abnormality in schizophrenia versus bipolar disorder: A whole brain cortical thickness, surface area, volume and gyrification analyses[J]. NeuroImage: Clinical, 2020, 25: 102131.

79. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5®)[M]. American Psychiatric Pub, 2013.

80. World Health Organization. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines[M]. World Health Organization, 1992.

81. Sheehan D V, Lecrubier Y, Sheehan K H, et al. The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10[J]. Journal of clinical psychiatry, 1998, 59(20): 22-33.

82. Gaser C, Dahnke R. CAT-a computational anatomy toolbox for the analysis of structural MRI data[J]. Hbm, 2016, 2016: 336-348.

83. Sled J G, Zijdenbos A P, Evans A C. A nonparametric method for automatic correction of intensity nonuniformity in MRI data[J]. IEEE transactions on medical imaging, 1998, 17(1): 87-97.

84. Ségonne F, Dale A M, Busa E, et al. A hybrid approach to the skull stripping problem in MRI[J]. Neuroimage, 2004, 22(3): 1060-1075.

85. Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

86. Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

87. Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

88. Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

89. Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001: 1189-1232.

90. Fischl B. FreeSurfer[J]. Neuroimage, 2012, 62(2): 774-781.

91. Desikan R S, Ségonne F, Fischl B, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest[J]. Neuroimage, 2006, 31(3): 968-980.

92. Fischl B, Salat D H, Busa E, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain[J]. Neuron, 2002, 33(3): 341-355.

93. Gillies R J, Kinahan P E, Hricak H. Radiomics: images are more than pictures, they are data[J]. Radiology, 2016, 278(2): 563-577.

94. Kaur T, Gandhi T K. Automated brain image classification based on VGG-16 and transfer learning[C] 2019 International Conference on Information Technology (ICIT). IEEE, 2019: 94-98.

95. Wang S H, Zhou Q, Yang M, et al. ADVIAN: Alzheimer's disease VGG-inspired attention network based on convolutional block attention module and multiple way data augmentation[J]. Frontiers in Aging Neuroscience, 2021, 13: 313.

96. Ajagbe S A, Amuda K A, Oladipupo M A, et al. Multi-classification of alzheimer disease on magnetic resonance images (MRI) using deep convolutional neural network (DCNN) approaches[J]. International Journal of Advanced Computer Research, 2021, 11(53): 51.

97. Ajagbe S A, Amuda K A, Oladipupo M A, et al. Multi-classification of alzheimer disease on magnetic resonance images (MRI) using deep convolutional neural

network (DCNN) approaches[J]. International Journal of Advanced Computer Research, 2021, 11(53): 51.

98. Yun J Y, Boedhoe P S W, Vriend C, et al. Brain structural covariance networks in obsessive-compulsive disorder: a graph analysis from the ENIGMA Consortium[J]. Brain, 2020, 143(2): 684-700.

99. Liu Z, Palaniyappan L, Wu X, et al. Resolving heterogeneity in schizophrenia through a novel systems approach to brain structure: individualized structural covariance network analysis[J]. Molecular Psychiatry, 2021: 1-13.

100. Wu H, Sun H, Wang C, et al. Abnormalities in the structural covariance of emotion regulation networks in major depressive disorder[J]. Journal of Psychiatric Research, 2017, 84: 237-242.

101. Kim S, Kim Y W, Jeon H, et al. Altered cortical thickness-based individualized structural covariance networks in patients with schizophrenia and bipolar disorder[J]. Journal of Clinical Medicine, 2020, 9(6): 1846.