OXFORD

Precision Clinical Medicine, 3(3), 2020, 202–213

doi: 10.1093/pcmedi/pbaa029
Advance Access Publication Date: 28 August 2020
Review

REVIEW

Recent advances of deep learning in psychiatric disorders

Lu Chen[1], Chunchao Xia[2] and Huaiqiang Sun[2,*]

[1]West China Medical Publishers, West China Hospital of Sichuan University, Chengdu 610041, China
[2]Department of Radiology, West China Hospital of Sichuan University, Chengdu 610041, China

*Correspondence: Huaiqiang Sun, sunhuaiqiang@scu.edu.cn

Abstract

Deep learning (DL) is a recently proposed subset of machine learning methods that has gained extensive attention in the academic world, breaking benchmark records in areas such as visual recognition and natural language processing. Different from conventional machine learning algorithm, DL is able to learn useful representations and features directly from raw data through hierarchical nonlinear transformations. Because of its ability to detect abstract and complex patterns, DL has been used in neuroimaging studies of psychiatric disorders, which are characterized by subtle and diffuse alterations. Here, we provide a brief review of recent advances and associated challenges in neuroimaging studies of DL applied to psychiatric disorders. The results of these studies indicate that DL could be a powerful tool in assisting the diagnosis of psychiatric diseases. We conclude our review by clarifying the main promises and challenges of DL application in psychiatric disorders, and possible directions for future research.

Key words: deep learning; machine learning; neuroimaging; autoencoders; convolutional neural networks; deep belief networks; mental disorders; psychiatric disorders

Introduction

Despite the increase in imaging studies on psychiatric disorders over the past decades,[1] the impact of imaging evidences on clinical practice remains limited. Most recently released DSM-5, the diagnostic manual for mental illness, does not incorporate the results of imaging studies.[2] The main reason for this is that most previous imaging studies adopted case-control comparison strategy, which compares the imaging features between patients with a disease of interest and individuals without, to see if statistically significant differences can be derived.[3,4] However, this research strategy has a number of limitations that hampered the translation of imaging findings to clinical applications. (1) Inter-group comparisons assume that the tested variables are normally distributed and the variance is consistent within each group, yet many studies show highly heterogeneity exists in the patient population.[5] (2) Case-control comparison can only reveal differences between groups, not allowing statistical inferences at the level of the individual.

Converting group differences into imaging biomarkers that can be used to assist diagnosis or prognosis requires estimating the inter-group and intra-group

Received: 3 August 2020; **Revised:** 24 August 2020; **Accepted:** 25 August 2020

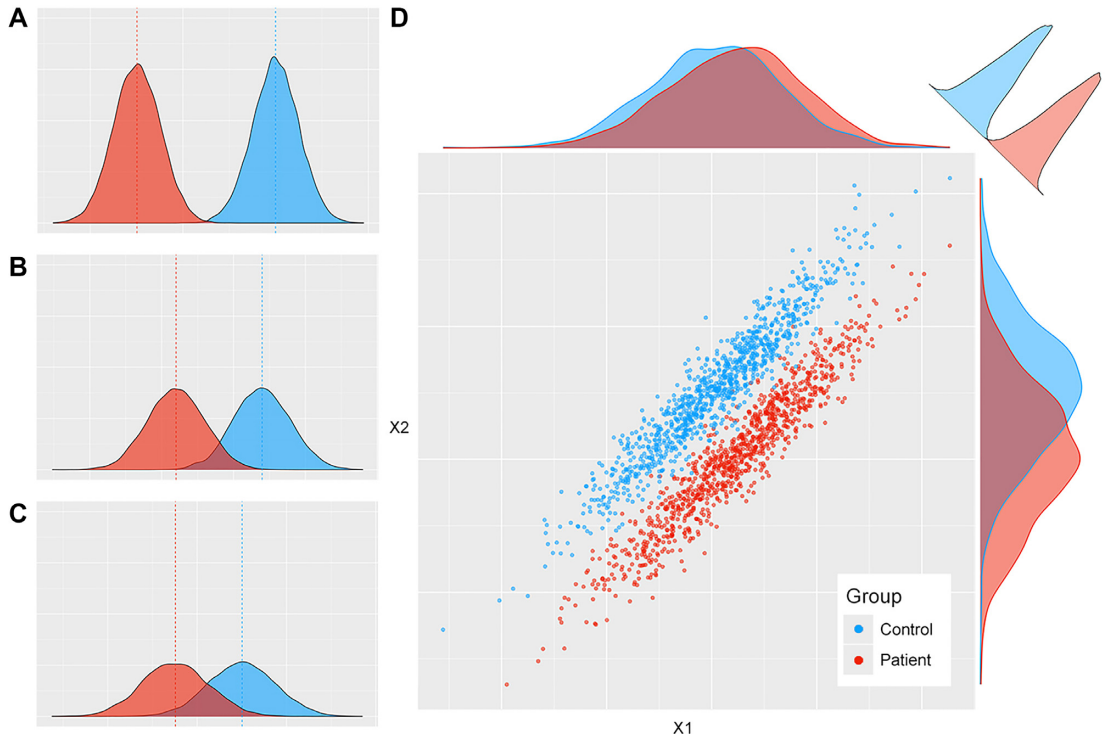

Figure 1. Three variables with varying degrees of inter-group variation (A, B, C). A possible joint distribution of two variables with small intra-group difference in their independent distribution (D).

variance of the imaging feature between the two groups. For example, three variables shown on the left side of Fig. 1 are all statistically different between groups, but only the variable shown in Fig. 1A is optimal for an imaging biomarker as intra-group difference is large and inter-group variance is small. In contrast, the variable shown in Fig. 1C with small intra-group difference and large inter-group variance will lead to poor diagnostic specificity and sensitivity if used as an imaging marker. Also case-control comparisons tend to be massive univariable statistical comparisons, thus leaving out the interaction between features. It may be the case that the distribution of one individual feature does not differ between two groups, but the joint distribution of two or more features does. As shown in Fig. 1D, the independent distributions of variables X1 and X2 are basically of no difference or slight difference between two groups, but a much better separation can be obtained if joint distribution of both variables is considered. All of these problems are not solved by traditional uni-variate case-control comparative research strategy.

Machine learning (ML) algorithms that have been widely used in email filtering, merchandise recommendation, and speech recognition are expected to solve the above mentioned problems. In general, ML can be classified into supervised learning and unsupervised learning. Supervised learning algorithms tend to summarize rules or patterns from already labeled data and form discriminatory models that can make predictions on new data; unsupervised learning algorithms explore possible structures in a data set based on the distribution of data points in unlabeled data. However, traditional ML was unable to work on raw image data and requires the use of expert design techniques to extract and construct informative features (a step known as "feature engineering"). In previous studies, researchers defined a variety of features from neuroimages and fed them into machine learning algorithms to construct disease classification or predictive models.[6]

The input features can be extracted from gray matter, such as cortical thickness[7] or gray matter density measured by brain mophometry[8] or white matter measured by diffusion MRI, such as anisotropy fraction (FA), mean diffusivity (MD).[9] Others have used brain network constructed from diffusion MRI or task/resting-state functional MRI or derived network parameters for disease discrimination models.[10–12] Unlike case-control comparative studies, the performance of machine learning studies is assessed primarily by their ability to predict new data samples, with commonly used metrics including accuracy, sensitivity, specificity, and area under the ROC curve (AUC).[13]

While feature engineer-based ML is still popular in the neuroimaging community, the recently proposed deep learning (DL)[14] has gained considerable attention in academia and industry. With significant improvement in various areas such as image classification, object detection, speech recognition and natural language processing, DL is superior to traditional machine learning in two aspects. First, compared with traditional machine learning algorithm, DL replaces artificial feature engineering with unsupervised or semi-supervised feature

learning and hierarchical feature extraction algorithms to identify the optimal representation automatically. This important capability overcomes the subjectivity in feature extraction and selection, especially in cases with extremely high feature dimension or when prior knowledge in feature selection is not conclusive. Another important characteristic of DL is the depth of models. By applying a hierarchy of non-linear transforms, DL is able to model very complex data patterns compared to traditional shallow models, which makes DL more suited for learning complicated patterns and subtle differences in data, especially the image data of human brain.

Overview

Due to limited ability to model complex data patterns and the need for complex and subjective feature engineer steps, conventional machine learning is experiencing bottlenecks in neuroimaging community. Given the advantages of DL,[14] researchers have begun to turn to this newly proposed method and have made numerous attempts at publicly available datasets.

The common workflow of DL in the application to neuroimaging involves data acquisition and labeling, model training and testing. First, the whole dataset is split into a training set and a test set according to a certain ratio (usually 5:1 to 4:1). The training set is used to optimize the weights of each node, aiming to capture characteristic pattern in the data. Next, the test set is fed into the trained model to evaluate whether the model can correctly predict the labels of the test set. In most neuroimaging studies, the available of sample size is limited, and model performance can be evaluated using a cross-validation approach in addition to the hold-out validation described above. In cross-validation, training and test run several times with different data partition scheme to get the average performance of the model. Information from each sample was fully utilized by participating both training and test phase. In the neuroimaging community there are currently two main types of DL application, depending on the type of input: (1) artificially constructed features, with the traditional ML algorithm being replaced by deep neural networks, such as deep belief network; (2) raw image data, fed into image-specific deep neural networks, such as convolutional neural network.

Commonly used architecture of deep neural networks

According to our review, we found the following three types of deep neural network and their variants to be the most widely used in the field of neuroimaging of psychiatric disorders.

Autoencoder

Due to the inherent high-dimensional nature of brain image data, a feature simplification approach is needed before model training. This is the essential work of autoencoder. Autoencoder is a specific deep neural network comprised of two main components.[15] One half is the encoder, which learns to generate low-dimensional representation of original input, while the other half is the decoder, which learns to use low-dimensional representations to reconstruct the data as close to the original input as possible (Fig. 2A). Meaningful features are extracted during the training process. With variations in loss function, variants of autoencoder can be obtained by changing the loss function, typically sparse autoencoders and denoising autoencoders.[16] The sparse autoencoder adds a regularization term in the loss function that constrains the node in hidden layer to output mostly zeros and a small number of non-zeros. The denoising autoencoder artificially adds noise to the input data and then tries to reconstruct the original input to improve the robustness to noise. Moreover, higher level of abstraction of original input can be learned by hierarchically stacking multiple autoencoders. The learned low-dimensional feature vector is then used as input for the classification layers, such as fully connected layers, for supervised learning.

Deep belief networks

Deep belief networks (DBNs) is the first model that successfully solve the optimization problem of deep neural networks.[17] The shallow neural network called restricted Boltzmann machine (RBM) is the building block of DBN.[18] A typical RBM is consist of a visible layer and a hidden layer, with full connections between the layers but no connections between the nodes within the layers. The hidden layer is trained to capture the stochastic representation from the visible layer. Several RBMs are concatenated together to form a DBN, where the hidden layer of the previous RBM is the visible layer of the next RBM, and the output of the previous RBM is the input of the next RBM (Fig. 2B). DBN is trained by a "layer-by-layer" approach, starting from the input layer by treating two adjacent layers of the network as an RBM for unsupervised training, after which the hidden layer of the previous RBM is treated as the visible layer of the next RBM. Once the layer-by-layer unsupervised training is complete, the model will be fine-tuned with back propagation training.

Convolutional neural networks

Early deep neural networks are not suitable for handling image or vision-related tasks, as spatial information is lost by stretching the image into a vector, and too many parameters are inefficient and difficult to train. Inspired by the human visual nervous system, convolutional neural networks (CNNs) are proposed for vision related learning task.[19] A CNN contains two special layers: convolutional layer and pooling layer (Fig. 2C). Within the convolutional layer, the convolutional kernels work like visual neurons, sliding sequentially across the image responding a restricted area in the image each time. Multiple convolution kernels are often used simultaneously to perceive different types of information in an image.

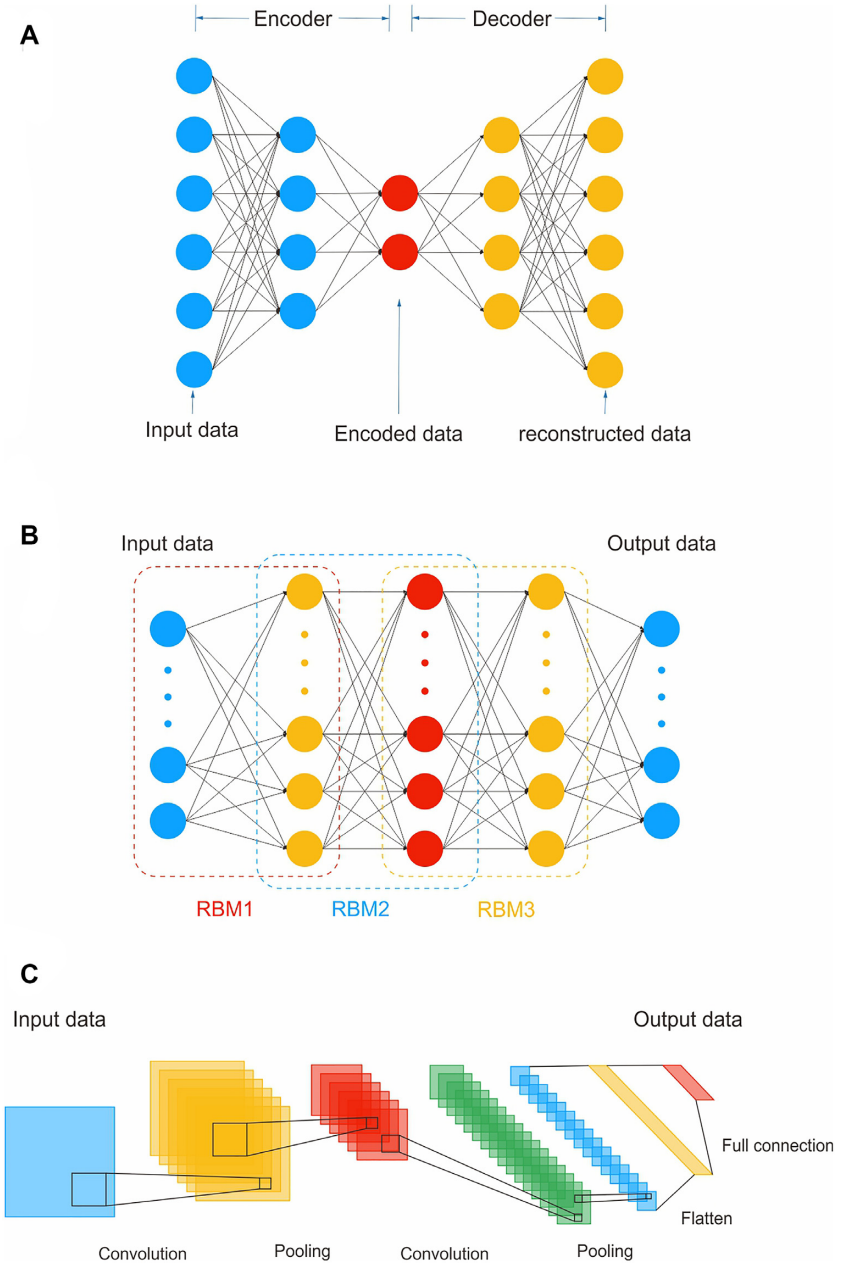

Figure 2. Commonly used architecture of deep neural Network: autoencoder (A), deep belief networks (B), and convolutional neural networks (C).

Weights in convolutional kernels are learned during the training process. The pooling comes after convolution. The essence of pooling is down sampling, which takes the input feature map and compresses it in some way. One of the more commonly used pooling processes is called max pooling which sends the maximum in the restricted area to the next level. With convolution and pooling operation, CNN gains the ability to effectively reduce large image to small image while preserving the characteristics of the image. More abstract features can be learned from image by alternating stacking of convolution and pooling layers.

DL studies of psychiatric disorders

In order to identify previous applications of DL in neuroimaging studies of psychiatric disorders, a search was conducted on 15 July 2020 across several databases (PubMed, IEEE Xplore, Science Citation Index, and Google Scholar) using the following search strategy: ("deep learning" OR "deep neural network" OR "convolutional neural network" OR "deep belief network") AND (psychiatry OR psychiatric OR "mental disorder" OR schizophrenia OR psychosis OR bipolar OR depression OR autism OR "Attention-deficit/hyperactive disorder" OR ADHD) AND

(neuroimaging OR MRI OR "Magnetic Resonance Imaging" OR fMRI OR "functional Magnetic Resonance Imaging"). This review mainly focuses on MRI research as it is a routine clinical examination without ionizing radiation. The initial search yielded a total of 54 articles. Then, we screened these articles for studies that had applied a deep learning method to neuroimaging data to investigate a psychiatric condition. We finally identified a total of 13 articles relevant to the current review.

All the identified articles are diagnostic studies to discriminate patients from healthy controls or discriminate between patient subtypes (Table 1). Most of them used a neuroimaging modality of resting-state functional MRI (rs-fMRI), followed by high-resolution structural MRI (sMRI); a few studies used a combination of the two modalities. The vast majority of studies were carried out in schizophrenia, ADHD and autism, diseases with challenge in diagnosis and public available dataset.

Schizophrenia

Schizophrenia is a chronic and serious mental disorder of unknown etiology with varying symptoms, including auditory and/or visual hallucination, disorganized speech or behavior. Some patients may experience cognitive impairment during the course of the illness. As there are no specific symptoms or clinical tests for schizophrenia, early diagnosis and intervention of schizophrenia are extremely difficult. Therefore, the objective diagnosis of schizophrenia through brain imaging has attracted considerable attention.

By utilizing structural MRI data, Pinaya *et al.* applied deep belief network (DBN) to features extracted from brain morphometry data for discriminating between healthy controls ($N = 83$) and patients with schizophrenia ($N = 143$).[20] The DBN achieved a classification accuracy of 73.6% within study cohort, but the accuracy reduced to 56.3% on an external data set consisted of 32 cases of first-episode psychosis, indicating that the patterns learned from patients with chronic schizophrenia and healthy controls were not suitable to classify patients with first episode psychosis. Therefore, it cannot be determined whether the features learned by deep neural networks are caused by disease or by the long-term medication. In another local structural MRI study, Latha *et al.* applied DBN only to ventricle region of a public data base which contains structural images from 72 patients with schizophrenia and 74 controls.[21] Finally, their proposed method achieved classification accuracy of 90% and area under ROC curve of 0.899.

Abnormalities of functional connectivity at resting state in patients with schizophrenia were reported in massive literatures, thus showing its potential as biomarker of clinical diagnosis. Han *et al.* applied autoencoder to resting-state connection matrix from 39 early-onset schizophrenia patients and 31 healthy controls.[22] The classification accuracy reached 79.3% (87.4% for sensitivity and 82.2% for specificity).

However, DL study on small sample is susceptible to over-fitting and poor generalizability. Zeng *et al.* applied a similar research strategy to a multi-site resting state fMRI data set from seven sites, which contains 357 schizophrenic patients and 377 controls.[23] Accuracy of 85% was achieved in multi-site pooling classification and 81% in leave-site-out transfer classification, respectively. The learned functional connectivity features reveal the dysregulation of the cortical-striatal-cerebellar circuit in patients with schizophrenia.

Most DL studies on schizophrenia used numerical features calculated from certain predefined procedures. This approach has the advantage of allowing easily control dimensionality of input features, but inevitably lead to the loss of potentially important information. Convolutional neural networks can automatically learn to discriminate features from image data. It can thus overcome the limitation of using human engineered numerical features as input. Qureshi *et al.* applied 3D CNN to maps of resting-state networks generated from independent component analysis (ICA) of public available 72 patients with schizophrenia and 72 healthy controls.[24] Their approach achieved a classification accuracy of $98.09\% \pm 1.01\%$ ten-fold cross-validated, and the AUC of 0.9982 ± 0.015.

Attention-deficit/hyperactive disorder

ADHD is a neurodevelopmental disorder characterized by age-inappropriate inattention, hyperactivity, and impulsivity. Due to the complexity of its pathological mechanism, there is a lack of objective diagnostic methods up to now. Imaging-based parameters may provide a useful objective adjunct to clinical psychiatric evaluation for diagnosing and subtyping ADHD. All four studies included here used imaging data from the ADHD-200 Consortium, a data-sharing design to understand the neural basis of ADHD.[25]

A fully connected deep neural network was applied to functional connectivity to identify children with ADHD from healthy controls.[26] The model successfully discriminated ADHD patients from healthy controls with an accuracy of 90%, while the two subtypes (ADHD-inattentive and ADHD-combined) were discriminated with an accuracy of 95%. The connection between frontal areas and the cerebellum is considered the most discriminating feature. To take full advantage of the functional and structural information in the ADHD-200 database, Zou *et al.* designed a multi-modality CNN architecture to combine fMRI and sMRI features and achieved an accuracy of 69.15%.[27] In contrast with artificially constructed features, Mao *et al.* constructed a spatio-temporal DL method called 4-D CNN based on granular computing which was trained based on derivative changes in entropy, and could be used to calculate the granularity at a coarse level by stacking layers.[28] The evaluations showed that the proposed method was superior to the

Table 1. Summary of reviewed studies.

Authors, year	Task	Sample size	Public dataset	Image modality	Input features	DL architecture	Performance
Pinaya, 2016	SZ vs HC	SZ:143 HC:83	No	sMRI	Morphometry	DBN	Accuracy: 73.6%
Latha, 2019	SZ vs HC	SZ:72 HC:74	No	sMRI	Morphometry	DBN	Accuracy: 90.0% AUC: 0.899
Han, 2017	SZ vs HC	SZ:39 HC:31	No	rs-fMRI	FC	Autoencoder	Accuracy: 90.0% Sensitivity: 87.4% Specificity:82.2%
Zeng, 2018	SZ vs HC	SZ:357 HC:377	COBRE (part)	rs-fMRI	FC		Accuracy: 85% (overall); 81% (leave site out)
Qureshi, 2019	SZ vs HC	SZ:72 HC:72	No	rs-fMRI	ICA maps	CNN	Accuracy: 98% AUC: 0.9982
Deshpande, 2015	ADHD vs HC ADHD-I vs ADHD-c	ADHD-I:173 ADHD-C:260 HC:744	ADHD-200	rs-fMRI	FC	FCN	Accuracy: 90% (ADHD vs HC); 95% (ADHD-I vs ADHD-C)
Zou, 2017	ADHD vs HC	ADHD:285 HC:491	ADHD-200	sMRI rs-fMRI	ReHo, ALFF, FC (rs-fMRI); tissue density (sMRI)	CNN	Accuracy: 69.15%
Mao, 2019	ADHD vs HC	HC:429 ADHD:359	ADHD-200	rs-fMRI	4D volumes	CNN + LSTM	Accuracy: 71.3% AUC: 0.80
Riaz, 2020	ADHD vs HC	ADHD:234 HC:232	ADHD-200	rs-fMRI	fMRI time series	FCN	Accuracy: 73.1% Sensitivity: 65.5% Specificity: 91.6%
Heinsfeld, 2018	ASD vs HC	ASD:505 HC:530	ABIDE	rs-fMRI	FC	Autoencoder	Accuracy: 70%
Guo, 2017	ASD vs HC	ASD:55 HC:55	ABIDE	rs-fMRI	FC	Autoencoder	Accuracy: 86.36%
Kong, 2019	ASD vs HC	ASD:78 HC:104	ABIDE	sMRI	SC	Autoencoder	Accuracy: 90.39% AUC: 0.9738
Akhavan, 2018	ASD vs HC HC:69	ASD:116	ABIDE	sMRI rs-fMRI	Mean intensity	DBN	Accuracy: 65.65% Sensitivity: 84% Specificity: 32.96%

SZ: schizophrenia.
HC: healthy controls.
sMRI: structural MRI.
DBN: deep belief network.
rs-fMRI: resting-state functional MRI.
ADHD: attention deficit/hyperactivity disorder.
FC: functional connectivity.
SC: structural connectivity.
FCN: fully connected network.
CNN: convolutional neural network.
ICA: independent component analysis.

traditional methods on the ADHD-200 dataset (accuracy: 71.3%, AUC: 0.80).

Novel architecture of deep neural network was designed specificly for ADHD-200 data. Riaz *et al.* proposed an end-to-end deep learning architecture to diagnose ADHD.[29] The model takes pre-processed fMRI time series as input and outputs a diagnosis, and is trained end-to-end using back-propagation and achieves classification accuracy of 73.1% (specificity 91.6%, sensitivity 65.5%). The model also suggests that the frontal lobe carries most discriminant power in classifying ADHD.

Although the abnormalities in patients are subtle, these studies show that DL can extract meaningful information from brain images to classify ADHD from controls, and more notably, to distinguish ADHD subtypes. However, it is worth noting that the samples in ADHD-200 are highly imbalanced, especially in terms of subtypes. With the exception of Riaz *et al.* who reported sensitivity and specificity,[29] all other studies reported model performance using the overall accuracy. This metric is simply the proportion of samples correctly classified, and therefore the classes imbalance is not considered. Given the highly imbalance in study population, the results reported in these studies may be exaggerated.

Autism spectrum disorder (ASD)

With regards to autism spectrum disorder, most DL studies carried on the publicly available Autism Brain Imaging Data Exchange (ABIDE) dataset.[30] Using resting-state fMRI data of 505 ADS patients and 530 matched controls from ABIDE, Heinsfel *et al.* applied autoencoder to flatted lower triangle of functional connectivity matrix generated using CC200 functional parcellation atlas.[31] Compared with the control group in the dataset, this study achieved 70% accuracy in diagnosing ASD versus control patients in the dataset. The patterns emerging from the classification show an inverse correlation of brain function between anterior and posterior areas of the brain, consistent with the evidence of anterior-posterior disruption in brain connectivity in ASD.[32]

Also using resting-state fMRI but from a single site (55 ASD patients and 55 TD controls) of ABIDE I, Guo *et al.* proposed deep neural network architecture with two stacked sparse autoencoder for feature extraction and selection as well as a softmax layer for classification.[33] In addition to the proposed DNN, a feature selection network based on stacked autoencoder was inserted before classification network. Results show that the best classification accuracy of 86.36% is generated by the DNN with feature selection network consisting of 3 hidden layers and 150 hidden nodes.

Some studies used structural image to construct the brain network as DL input. In the study carried by Kong *et al.*, the T1w images from ABIDE I were segmented by FreeSurfer software. The gray matter volume of each segmented region is defined as the node of the network, the differences between each pair of segmented regions is defined as the edge. Then, the "edges" were ranked and the first 3000 were fed into the same DNN proposed by Guo *et al.* The method they proposed has an accuracy of 90.39% and the AUC of 0.9738 for ASD/HC classification.[34] However, the results are over-optimistic as the feature selection was conducted on the whole dataset, which would cause information leak from test data to the training process.[35]

Another study combined both structural and functional MRI to discriminate ASD in young children. With 185 individuals (116 ASD and 69 HC), aged between 5 and 10 years old from ABIDE, the best combination comprised rs-fMRI, GM, and WM for DBN of depth 3 with 65.56% accuracy (sensitivity = 84%, specificity = 32.96%, F1 score = 74.76%) obtained via 10-fold cross-validation.[36] However, the result may be unreliable as the features are mean intensity of weighted images, which may be affected by many factors, such as field strength, coil configuration even the weight of participant.

Challenge

Small sample size and risk of overfitting

Deep learning models have millions of weights to be learned during training phase, thus require a large amount of samples to learn complex patterns compared to traditional machine learning methods. But how many samples are meaningful for DL in neuroimaging studies remains inconclusive, possibly due to the limited studies currently available. If a deep network is trained on very limited samples, especially image data with high dimension, it is possible that the trained model works perfectly on training set but poorly in test set, a problem known as "overfitting".[37] The best solution to overfitting is collecting more training samples. Large image datasets of psychiatric disorder are not easy to obtain for reasons. Firstly, high-resolution structural imaging and functional MRI are not included in routine clinical scans. The purpose of routine clinical scans is to check whether there are organic lesions in patients with psychiatric disorder. Limited in modality and resolution, the images from routine scan may not carry sufficient information for disease discrimination. Hence, prospective studies with designed scanning protocol can only provide small datasets. Secondly, sharing medical data has become increasingly difficult over the past few years due to increasingly strict laws on patient privacy. Institutions therefore have to take a longer time to accumulate enough data for applying DL analysis.

Overfitting can also be avoided by using regularization methods such as Drop-out, which is achieved by modifying the neural network structure in real time during training.[38] Each node of the neural network is given a probability to be temporarily ignored during training. In each training epoch, the ignored nodes are randomly selected based on the preset probability. As a result, multiple smaller neural networks are actually trained. This

mechanism will ensure that the neural network does not "over-match" the training samples, which will help mitigate the overfitting problem. Another approach that can mitigate overfitting is data augmentation, which can generate more equivalent data based on limited real data according to certain transform.[39] For conventional vision task, geometric transformations, including flipping, rotating, cropping and scaling, are the main implementation of data augmentation. However, it has been reported that geometric transformation is not suitable for medical imaging.[40] Several studies have attempted to use generative adversarial network based methods for medical image augmentation and have been successfully applied to skin lesions[41] and liver cancer.[42] However, whether this method is applicable to brain images of psychiatric disorders remains to be investigated. While data augmentation does not add substantive information, it allows the trained deep neural network to be robustness to body positioning and individual differences. Transfer learning is also a way to mitigate small sample size problem.[43] Two stages of model training are involved in transfer learning. First the model was pre-trained on a large-scale benchmark dataset, and then fine-tuned on a small but study specific dataset. However, the application of transfer learning in medical imaging remains limited, mainly because there is currently no widely recognized benchmark dataset for medical imaging.

Lack of standardization in data acquisition and uneven data quality

Most studies in current review used public available datasets, such as ADIBE and ADHD-200. Although the sample size in these datasets is relatively large, the data are quite heterogeneous as they are from multiple sites. Each site is different in terms of field strength, coil configuration and imaging sequences. It has been proved that the performance of machine learning models can be affected by imaging parameters. Moreover, echo planar imaging (EPI) sequence, which is widely used in diffusion and functional MRI, tends to be sensitive to magnetic field inhomogeneity and highly susceptible to image distortion and signal loss at the junction of tissues and air, such as the frontal and temporal lobes,[44] which are the very brain regions that psychiatric research focuses on. Although many correction methods have been proposed in the field of MR physics and medical imaging,[45] the need for additional scans and complex calculations has prevented these corrective methods from being widely used in clinical oriented research. In addition, the analysis of diffusion MR data has special requirements for image acquisition, and many previous studies have used diffusion tensor imaging (DTI) technique to investigate the structural connectivity in patients with psychiatric disorders.[46] But the limitation of the tensor model is the inability to resolve fiber crossing,[47] so the connectom constructed based on tensor model may miss a large number of possible connections. Meanwhile, it has been suggested that the functional connectivity not only based on correlations of low-frequency signals but also on high-frequency correlations.[48] The resting-state fMRI acquisition protocol with repetition time (TR) equal or greater than 2 seconds used in most previous studies cannot reconstruct such high-frequency functional connectivity.

Perspective

In view of the above-mentioned challenges and pitfalls, the future development of imaging studies of psychiatric disorders can be carried out in the following directions.

Establishment of standard image database of psychiatric disorders

The image is not only the picture but also the data that can be mined.[49] The establishment of a high quality database is an important work for the further development of DL based imaging research of psychiatric disorders. A predictive model with satisfactory performance must be built on high quality data, so standardized data acquisition protocols and databases are the primary guarantee for the translation of imaging research to clinic practice. Researchers all over the world have worked extensively to establish standard MRI brain image acquisition protocols, the most influential one is the protocols recommended by the Human Connectome Project (HCP).[50] The HCP is the state-of-art large-scale research program funded by the National Institutes of Health (NIH) with the participation of more than 100 researchers from 10 research institutions. This project aims to understand the principles of how the human brain works via structural and functional connectivity. The HCP recommends scanning protocols including structural diffusion MRI, resting-state fMRI, task fMRI, as well as a set of tools for processing the raw images, like tissue segmentation, distortion correction and model fitting.[51,52]

The HCP recommends high resolution anatomical T2w scans, using the T1w/T2w ratio to map myelin content across the cortical surface, thereby non-invasively distinguishing many architectonic areas (Fig. 3A, B, C). For diffusion MR, HCP recommends to perform multiple diffusion weighted directions and multiple shells acquisition with reversed phase encoding directions for both distortion correction[45] and constrained spherical deconvolution model fitting, which is able to resolve "fiber crossing" (Fig. 3D, E).[53] Also, diffusion acquisition can be accelerated by deep learning methods to significantly reduce the total scan time.[54,55] With the advance of pulse sequence, the multi-band technique can significantly increase the spatial and temporal resolution of functional MRI.[56] HCP recommends to use this technique to capture high-frequency connectivity in the brain. In addition, for multi-center studies, data harmonization across sites and devices should be performed to remove non-diseases related confounders before pooling together.[57,58]

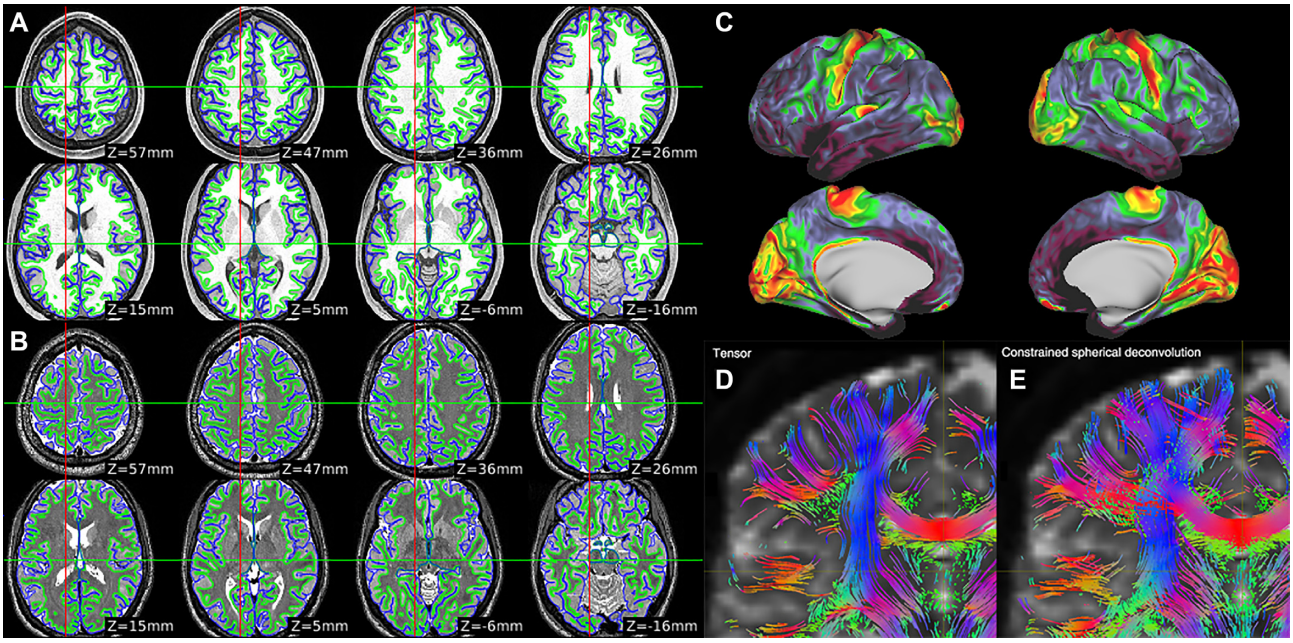

Figure 3. Images acquired by HCP recommended protocols. High-resolution T1 weighted (A) and T2 weighted (B) volume. Myelin map generated by T1w/T2w ratio (C). Fiber tractography generated from tensor (D) and constrained spherical deconvolution (E) model, "fiber crossing" can be solved by CSD model correctly.

However, these recommended scanning protocols are designed for healthy individuals with considerable long scan time. Patients with psychiatric disorder may have a much lower tolerance level than healthy people, and therefore the scanning protocol needs to be simplified or set break point at appropriate locus, while ensuring that sufficient information for analysis is captured.

Automated deep learning (autoDL)

The development and implementation of DL methodology into medical imaging research still faces a main challenge. Although there are well-established deep learning frameworks, such as TensorFlow and Pytorch, applying these frameworks still requires researchers to input code to call functions in these libraries. We also find that most studies included in this review were done by computer experts but not clinical researchers. The successful application of deep neural network to a given problem crucially relies on artificial intervention in many steps, such as data preparation, architecture selection, parameter tuning. Especially, the optimal architecture of deep neural network is not estimated as part of the learning process but is defined as a priori. As the complexity of these tasks is often beyond non-DL-experts, autoDL was recently proposed in the field of computer science that has the potential to help non-experts use deep learning off-the-shelf.[59]

For classification tasks, autoDL automatically matches generic neural network architectures with a given imaging dataset, fine-tunes the network with the goal of optimizing discriminative performance, and creates a prediction model as output. Clinical researchers without coding experience can quickly implement an appropriate DL workflow to their data and estimate the performance of the generated model. With few medical imaging studies using this technique, autoDL is still an active research field with no fixed rules currently.

Graph/geometry deep learning

DL has been successfully applied to several types of input, like feature vectors or images, however, all these data are Euclidean data. In the field of neuroimaging research, there are also a large amount of non-Euclidean data. The most common non-Euclidean data are connectom and brain surface. Connectom is essentially a graph. In most of the previous studies, connectom was first converted into binary network by predefined thresholds, then a series of characteristic parameters were calculated as the subsequent input features according to the graph theory. This kind of feature construction, which relies on manual intervention, will inevitably result in information loss and bias.

The brain surface is indeed a graphic, an extension of the graph, consisting of a large number of polygonal meshes. Compared with image volume, mesh is a more superior data structure to represent the shape of objects without the interference of noise in images (Fig. 4). The previous brain surface based analysis only performed univariate statistical analysis at each vertex of the mesh, which did not fully exploit the advantages of mesh data in shape representation.[60]

The previous DL techniques are unable to input non-Euclidean space data directly. Recently, a substantial amount of research has been devoted to the application of DL methods to graphs, resulting in beneficial advances

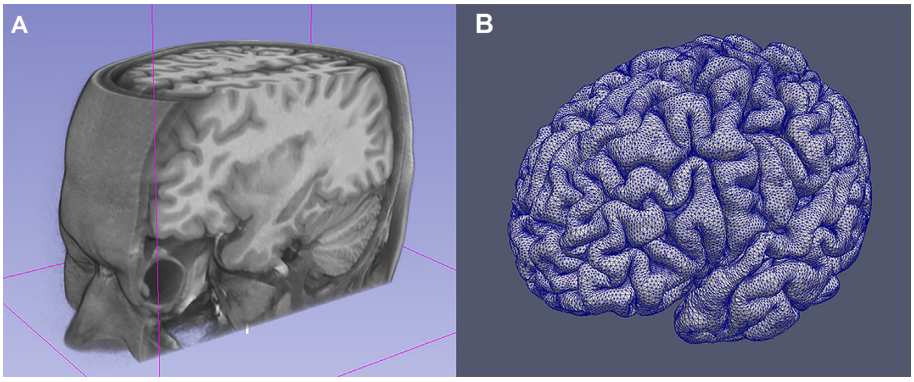

Figure 4. Image volume (A) and mesh of brain surface (B).

in graph deep neural network.[61] This new technology successfully redefines important operations such as convolution and pooling in non-Euclidean space, so that the graph or mesh can be used directly as the input and output the prediction results, avoiding loss of information in the data preparation stage to the greatest possible extent. Although this technique has not been widely implemented in the field of neuroimaging, with the support of large datasets, it is expected to replace the conventional voxel-based morphometry and graph theory analysis in classification research.

Summary

Although DL techniques have been explored extensively in various aspects of medical imaging, they are still in a relatively early stage, and most applications are still simple two- or three-classification problems. In the clinical practice of psychiatric disorders, clinicians are often faced with more complex situations. Therefore, for a long time in the future, DL cannot replace the physician's position in diagnosis or treatment decision-making. But the combination of imaging examination and DL will gradually develop into a laboratory test to assist the diagnosis of psychiatric disorders, providing clinicians with a solid basis for precise and efficient diagnosis.

Despite many challenges, the rapid development of DL algorithms and computer hardware, the establishment of standardized medical image databases, and the formation of multi-center data sharing mechanisms, make it possible to study more complex clinical problems and obtain models with better generalization performance. The combined development of psychiatric imaging and machine learning will be the trend, and will become an indispensable tool for clinical diagnosis and treatment of psychiatric diseases in the future.

Author contributions

L.C. drafted the manuscript; C.C.X. edited the manuscript; H.Q.S. supervised and revised the manuscript.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 91859203) and Young Elite Scientists Sponsorship Program by CAST (YESS20160060).

Conflict of interest

None declared.

References

1. Mitelman SA. Transdiagnostic neuroimaging in psychiatry: A review. *Psychiatry Res* 2019;**277**:23–38. doi:10.1016/j.psychres.2019.01.026.
2. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association; 2013. doi:10.1176/appi.books.9780890425596.
3. Shenton ME, Dickey CC, Frumin M, *et al*. A review of MRI findings in schizophrenia. *Schizophr Res* 2001;**49**:1–52. doi:10.1016/S0920-9964(01)00163-3.
4. Brown GG, Eyler LT. Methodological and conceptual issues in functional magnetic resonance imaging: Applications to schizophrenia research. *Annu Rev Clin Psychol* 2006;**2**:51–81. doi:10.1146/annurev.clinpsy.2.022305.095241.
5. Wardenaar KJ, de Jonge P. Diagnostic heterogeneity in psychiatry: Towards an empirical solution. *BMC Med* 2013;**11**:201. doi:10.1186/1741-7015-11-201.
6. Walter M, Alizadeh S, Jamalabadi H, *et al*. Translational machine learning for psychiatric neuroimaging. *Prog Neuro-Psychopharmacology Biol Psychiatry* 2019;**91**:113–21. doi:10.1016/j.pnpbp.2018.09.014.
7. Fischl B, Dale AM. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc Natl Acad Sci* 2000;**97**:11050–5. doi:10.1073/pnas.200033797.
8. Ashburner J, Friston KJ. Voxel-based morphometry—the methods. *Neuroimage* 2000;**11**:805–21. doi:10.1006/nimg.2000.0582.
9. Vergara VM, Mayer AR, Damaraju E, *et al*. Detection of mild traumatic brain injury by machine learning classification using resting state functional network connectivity and fractional anisotropy. *J Neurotrauma* 2017;**34**:1045–53. doi:10.1089/neu.2016.4526.
10. Iturria-Medina Y. Anatomical Brain Networks on the prediction of abnormal brain states. *Brain Connect* 2013;**3**:1–21. doi:10.1089/brain.2012.0122.

11. Rubinov M, Sporns O. Complex network measures of brain connectivity: Uses and interpretations. *Neuroimage* 2010;**52**:1059–1069. doi:10.1016/j.neuroimage.2009.10.003.
12. Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fMRI: A tutorial overview. *Neuroimage* 2009;**45**:S199–209. doi:10.1016/j.neuroimage.2008.11.007.
13. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit* 1997;**30**:1145–59. doi:10.1016/S0031-3203(96)00142-2.
14. Kriegeskorte N. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annu Rev Vis Sci* 2015;**1**:417–46. doi:10.1146/annurev-vision-082114-035447.
15. Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J* 1991;**37**:233–43. doi:10.1002/aic.690370209.
16. Dong G, Liao G, Liu H, *et al*. A review of the autoencoder and its variants: A comparative perspective from target recognition in synthetic-aperture radar images. *IEEE Geosci Remote Sens Mag* 2018;**6**:44–68. doi:10.1109/MGRS.2018.2853555.
17. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. *Neural Comput* 2006;**18**:1527–54. doi:10.1162/neco.2006.18.7.1527.
18. Hjelm RD, Calhoun VD, Salakhutdinov R, *et al*. Restricted Boltzmann machines for neuroimaging: An application in identifying intrinsic networks. *Neuroimage* 2014;**96**:245–60. doi:10.1016/j.neuroimage.2014.03.048.
19. Valueva MV, Nagornov NN, Lyakhov PA, *et al*. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Math Comput Simul*. 2020;**177**:232–43. doi:10.1016/j.matcom.2020.04.031.
20. Pinaya WHL, Gadelha A, Doyle OM, *et al*. Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Sci Rep* 2016;**6**:38897. doi:10.1038/srep38897.
21. Latha M, Kavitha G. Detection of Schizophrenia in brain MR images based on segmented ventricle region and deep belief networks. *Neural Comput Appl* 2019;**31**:5195–206. doi:10.1007/s00521-018-3360-1.
22. Han S, Huang W, Zhang Y, *et al*. Recognition of early-onset schizophrenia using deep-learning method. *Appl Informatics* 2017;**4**:16. doi:10.1186/s40535-017-0044-3.
23. Zeng L-L, Wang H, Hu P, *et al*. Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. *EBioMedicine* 2018;**30**:74–85. doi:10.1016/j.ebiom.2018.03.017.
24. Qureshi MNI, Oh J, Lee B. 3D-CNN based discrimination of schizophrenia using resting-state fMRI. *Artif Intell Med* 2019;**98**:10–17. doi:10.1016/j.artmed.2019.06.003.
25. Consortium T. The ADHD-200 Consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. *Front Syst Neurosci* 2012;**6**:62. doi:10.3389/fnsys.2012.00062.
26. Deshpande G, Wang P, Rangaprakash D, *et al*. Fully connected cascade artificial neural network architecture for attention deficit hyperactivity disorder classification from functional magnetic resonance imaging data. *IEEE Trans Cybern* 2015;**45**:2668–79. doi:10.1109/TCYB.2014.2379621.
27. Zou L, Zheng J, Miao C, *et al*. 3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI. *IEEE Access* 2017;**5**:23626–36. doi:10.1109/ACCESS.2017.2762703.
28. Mao Z, Su Y, Xu G *et al*. Spatio-temporal deep learning method for ADHD fMRI classification. *Inf Sci (Ny)* 2019;**499**:1–11. doi:10.1016/j.ins.2019.05.043.
29. Riaz A, Asad M, Alonso E *et al*. DeepFMRI: End-to-end deep learning for functional connectivity and classification of ADHD using fMRI. *J Neurosci Methods* 2020;**335**:108506. doi:10.1016/j.jneumeth.2019.108506.
30. Di Martino A, Yan C-G, Li Q, *et al*. The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol Psychiatry* 2014;**19**:659–67. doi:10.1038/mp.2013.78.
31. Heinsfeld AS, Franco AR, Craddock RC, *et al*. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage Clin* 2018;**17**:16–23. doi:10.1016/j.nicl.2017.08.017.
32. Dajani DR, Uddin LQ. Local brain connectivity across development in autism spectrum disorder: A cross-sectional investigation. *Autism Res* 2016;**9**:43–54. doi:10.1002/aur.1494.
33. Guo X, Dominick KC, Minai AA, *et al*. Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. *Front Neurosci* 2017;**11**:460. doi:10.3389/fnins.2017.00460.
34. Kong Y, Gao J, Xu Y, *et al*. Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. *Neurocomputing* 2019;**324**:63–8. doi:10.1016/j.neucom.2018.04.080.
35. Arbabshirani MR, Plis S, Sui J, *et al*. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage* 2017;**145**:137–65. doi:10.1016/j.neuroimage.2016.02.079.
36. Akhavan Aghdam M, Sharifi A, Pedram MM. Combination of rs-fMRI and sMRI Data to discriminate autism spectrum disorders in young children using deep belief network. *J Digit Imaging* 2018;**31**:895–903. doi:10.1007/s10278-018-0093-8.
37. Whelan R, Garavan H. When optimism hurts: Inflated predictions in psychiatric neuroimaging. *Biol Psychiatry* 2014;**75**:746–8. doi:10.1016/j.biopsych.2013.05.014.
38. Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;**15**:1929–58.
39. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. December 2017. arXiv:1712.04621.
40. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data* 2019;**6**:60. doi:10.1186/s40537-019-0197-0.
41. Baur C, Albarqouni S, Navab N. MelanoGANs: High Resolution Skin Lesion Synthesis with GANs. April 2018. arXiv:1804.04338.
42. Frid-Adar M, Klang E, Amitai M, *et al*. Synthetic data augmentation using GAN for improved liver lesion classification.In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018:289–293. doi:10.1109/ISBI.2018.8363576.
43. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med Image Anal* 2019;**54**:280–96. doi:10.1016/j.media.2019.03.009.
44. Jezzard P, Balaban RS. Correction for geometric distortion in echo planar images from B0 field variations. *Magn Reson Med* 1995;**34**:65–73. doi:10.1002/mrm.1910340111.

45. Andersson JLR, Skare S, Ashburner J. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage* 2003;**20**:870–88. doi:10.1016/S1053-8119(03)00336-7.
46. O'Donoghue S, Holleran L, Cannon DM, *et al*. Anatomical dysconnectivity in bipolar disorder compared with schizophrenia: A selective review of structural network analyses using diffusion MRI. *J Affect Disord* 2017;**209**:217–28. doi:10.1016/j.jad.2016.11.015.
47. Jeurissen B, Descoteaux M, Mori S, *et al*. Diffusion MRI Fiber tractography of the brain. *NMR Biomed* 2019;**32**:e3785. doi:10.1002/nbm.3785.
48. Trapp C, Vakamudi K, Posse S. On the detection of high frequency correlations in resting state fMRI. *Neuroimage* 2018;**164**:202–13. doi:10.1016/j.neuroimage.2017.01.059.
49. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data. *Radiology* 2016;**278**:563–77. doi:10.1148/radiol.2015151169.
50. Van Essen DC, Smith SM, Barch DM, *et al*. The WU-Minn Human Connectome Project: An overview. *Neuroimage* 2013;**80**:62–79. doi:10.1016/j.neuroimage.2013.05.041.
51. Van Essen DC, Ugurbil K, Auerbach E, *et al*. The Human Connectome Project: A data acquisition perspective. *Neuroimage* 2012;**62**:2222–31. doi:10.1016/j.neuroimage.2012.02.018.
52. Glasser MF, Sotiropoulos SN, Wilson JA, *et al*. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* 2013;**80**:105–24. doi:10.1016/j.neuroimage.2013.04.127.
53. Tournier J-D, Yeh C-H, Calamante F, *et al*. Resolving crossing fibres using constrained spherical deconvolution: Validation using diffusion-weighted imaging phantom data. *Neuroimage* 2008;**42**:617–25. doi:10.1016/j.neuroimage.2008.05.002.
54. Li Z, Gong T, Lin Z, *et al*. Fast and robust diffusion kurtosis parametric mapping using a three-dimensional convolutional neural network. *IEEE Access* 2019;**7**:71398–411. doi:10.1109/ACCESS.2019.2919241.
55. Lin Z, Gong T, Wang K, *et al*. Fast learning of fiber orientation distribution function for MR tractography using convolutional neural network. *Med Phys* 2019;**46**:3101–16. doi:10.1002/mp.13555.
56. Barth M, Breuer F, Koopmans PJ, *et al*. Simultaneous multislice (SMS) imaging techniques. *Magn Reson Med* 2016;**75**:63–81. doi:10.1002/mrm.25897.
57. Tong Q, Gong T, He H, *et al*. A deep learning–based method for improving reliability of multicenter diffusion kurtosis imaging with varied acquisition protocols. *Magn Reson Imaging* 2020. doi:10.1016/j.mri.2020.08.001.
58. Fortin JP, Cullen N, Sheline YI, *et al*. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* 2018;**167**:104–120. doi:10.1016/j.neuroimage.2017.11.024.
59. Faes L, Wagner SK, Fu DJ, *et al*. Automated deep learning design for medical image classification by healthcare professionals with no coding experience: A feasibility study. *Lancet Digit Heal* 2019;**1**:e232–42. doi:10.1016/S2589-7500(19)30108-6.
60. Dale AM, Fischl B, Sereno MI. Cortical Surface-Based Analysis. *Neuroimage* 1999;**9**:179–94. doi:10.1006/nimg.1998.0395.
61. Henaff M, Bruna J, LeCun Y. Deep convolutional networks on graph-structured data. June 2015. arXiv:1506.05163.