Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

Going deep into schizophrenia with artificial intelligence

Jose A. Cortes-Briones[a,b,c,*,1], Nicolas I. Tapia-Rivas[d,1], Deepak Cyril D'Souza[a,b,c], Pablo A. Estevez[d]

[a] Schizophrenia and Neuropharmacology Research Group, VA Connecticut Healthcare System, West Haven, CT, USA
[b] Abraham Ribicoff Research Facilities, Connecticut Mental Health Center, New Haven, CT, USA
[c] Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA
[d] Department of Electrical Engineering, University of Chile, Santiago, Chile

ARTICLE INFO

Keywords:
Schizophrenia
Deep learning
Artificial intelligence
Machine learning
Psychosis
Prediction

ABSTRACT

Despite years of research, the mechanisms governing the onset, relapse, symptomatology, and treatment of schizophrenia (SZ) remain elusive. The lack of appropriate analytic tools to deal with the heterogeneity and complexity of SZ may be one of the reasons behind this situation. Deep learning, a subfield of artificial intelligence (AI) inspired by the nervous system, has recently provided an accessible way of modeling and analyzing complex, high-dimensional, nonlinear systems. The unprecedented accuracy of deep learning algorithms in classification and prediction tasks has revolutionized a wide range of scientific fields and is rapidly permeating SZ research. Deep learning has the potential of becoming a valuable aid for clinicians in the prediction, diagnosis, and treatment of SZ, especially in combination with principles from Bayesian statistics. Furthermore, deep learning could become a powerful tool for uncovering the mechanisms underlying SZ thanks to a growing number of techniques designed for improving model interpretability and causal reasoning. The purpose of this article is to introduce SZ researchers to the field of deep learning and review its latest applications in SZ research. In general, existing studies have yielded impressive results in classification and outcome prediction tasks. However, methodological concerns related to the assessment of model performance in several studies, the widespread use of small training datasets, and the little clinical value of some models suggest that some of these results should be taken with caution.

1. Introduction

Schizophrenia (SZ) is a heterogeneous and complex disorder, aptly coined the "group of schizophrenias" by Bleuler in 1911 (Bleuler, 1950). SZ is perhaps best understood as an umbrella term, comprising several distinct disorders with partially overlapping phenomenology and neural correlates (Alnæs et al., 2019; Brugger and Howes, 2017; Farmer et al., 1983). Furthermore, SZ seems to result from the complex interactions of many endogenous and exogenous factors shaping neurodevelopment (Alnæs et al., 2019; Bowen et al., 2019; Brugger and Howes, 2017; Guest et al., 2013; Liang and Greenwood, 2015; Voineskos, 2015). Thus, it is becoming increasingly clear that it is unlikely, and overly simplistic, that single-cause mechanisms or simple (linear) relationships between small sets of biomarkers could explain or even classify the different forms of SZ. The lack of appropriate analytic tools to address this heterogeneity and complexity is likely one of the reasons why, after years of research using traditional statistical and machine learning (ML) approaches, the mechanisms governing the onset, relapse, phenomenology, cognitive deficits, and treatment of SZ remain elusive. However, during the past five years, a collection of powerful techniques, algorithms, and ideas coming from a subfield of ML called *deep learning* has infused the field of neuropsychiatry with an accessible way of modeling and analyzing complex, high-dimensional, nonlinear problems.

Artificial intelligence (AI) is the science and engineering of making computers (machines) solve problems and behave in ways generally considered to be intelligent or, until recently, unique to human intelligence (McCarthy, 1983). ML is one of the many branches of AI and it refers to a family of methods for solving problems such as classification, prediction, and system modeling. Differently from standard computer programs, ML algorithms *learn* to solve problems by *training* over large numbers of examples through an iterative optimization process (e.g., error minimization). Deep learning is a type of ML that uses deep neural

* Corresponding author at: VA Connecticut Healthcare System, 950 Campbell Avenue, Building 1, Suite 9140, West Haven, CT 06516, USA.
E-mail address: jose.briones@yale.edu (J.A. Cortes-Briones).

[1] These authors contributed equally to this manuscript.

https://doi.org/10.1016/j.schres.2021.05.018
Received 14 December 2020; Received in revised form 24 May 2021; Accepted 27 May 2021
Available online 5 June 2021

networks (DNNs), algorithms inspired by the nervous system consisting of multiple interconnected layers of nonlinear processing units called artificial neurons. Similar to the nervous system, DNN learning consists of changes in the strength (weights) of the connections between neurons that occur during training (LeCun et al., 2015).

One of the main differentiating factors of deep learning approaches is the type of input data that they work on. Traditional ML (e.g., support vector machines [SVMs], decision trees, and random forests) and statistical (e.g., multivariate regressions) approaches operate over highly processed data, i.e., they utilize a limited number of features *extracted from* high-dimensional raw data under the theory-motivated assumption that they are relevant for a task (e.g., power ratios between electroencephalography [EEG] frequency bands). In contrast, DNNs can work *directly on* high-dimensional raw or minimally processed data and learn to extract the *best* features for a task automatically. This difference is believed to underlie part of the high performance showed by deep learning compared to traditional ML on problems involving complex high-dimensional data and nonlinear input/output relationships, including mapping high-dimensional unstructured data to categories (e.g., diagnosing SZ using simultaneous EEG/functional magnetic resonance imaging [fMRI] multimodal data) or predicting the future behavior of nonlinear dynamical systems (e.g., patient prognosis from electrophysiological data).

While the theoretical background for DNNs has been around for decades (LeCun et al., 1998; Rumelhart et al., 1986), cheap access to high-powered processors (GPUs), development of large electronic datasets, and algorithmic improvements in the past 10 years led to the recent widespread adoption and development of DNN technology. This expansion has afforded results that seemed impossible just a few years ago. For example, DNNs are capable of translating brain activity directly into language (Anumanchipalli et al., 2019), beating humans in the ancient game of Go (Silver et al., 2016), and outperforming experts in the detection of melanoma in dermoscopic images (Brinker et al., 2019). The field of psychiatry has been slow to adopt DNNs, but it is catching up and delivering promising results, including the classification of SZ patients from neuroimages (Kim et al., 2016) and speech patterns (Naderi et al., 2019), among others.

While it is easy to see how DNNs may benefit diagnosis, subtyping, and treatment decisions in SZ, it is less clear how DNNs could improve our understanding of the mechanisms underlying the disorder. DNNs are commonly considered “black box” models, meaning that the series of nonlinear transformations and high-dimensional representations used by DNNs to solve problems are too complicated to be interpreted. Thus, differently from regression models in which the model's coefficients (β) can provide easy-to-interpret information on the relationship (direction, strength) between the inputs and outputs of a system (e.g., traumatic event/PTSD), DNN models would not provide any insight into the internal mechanism of problems or systems. This hard notion of “black box” is being increasingly softened by a growing number of explainable AI techniques designed to reveal the inputs that were most relevant for generating the output, after the DNN has made a prediction. Furthermore, researchers have started working on DNNs equipped with cause-and-effect models (hypotheses) of reality (problems), that eventually may lead to novel mechanistic explanations of SZ.

The purpose of this article is to introduce the SZ researcher to the possibilities that DNNs offer to the field. We will review some of the latest developments in DNNs used for diagnosing and predicting clinical outcomes in SZ, and we will go over some of the techniques that could be used to better understand the neural mechanisms of the disorder. Finally, we will briefly review and discuss some of the current trends in AI including Bayesian and causal models, examine ethical issues associated with using deep learning for SZ research, and outline some future directions in the field of AI-powered SZ research.

2. Deep learning

2.1. General overview

Deep learning refers to the use of deep (multi-layer) artificial neural networks (DNNs), a family of problem-solving or system-modeling algorithms inspired by the nervous system that, differently from traditional software, learns how to solve problems through training (LeCun et al., 2015). DNNs have excelled in tasks that, until recently, were thought to be the exclusive domain of human expertise. For example, DNNs have surpassed human performance in complex tasks such as the ancient game of Go (Silver et al., 2016) and have demonstrated excellent performance in speech recognition (Baevski et al., 2020), language translation (Arivazhagan et al., 2019), text understanding and generation (Brown et al., 2020a), and object detection and recognition in images and videos (Liu et al., 2020), among others.

The basic element of any artificial neural network (ANN) is the artificial neuron, a mathematical model inspired by the input integration and nonlinear activation (output) of biological neurons (Fig. 1). In general, an artificial neuron consists of a weighted sum of numerical inputs followed by a nonlinear differentiable (smooth) activation function that transforms the result of the sum into the neuron's numerical output (activation). The weights multiplying the inputs are the learnable parameters of ANNs, and they model the synaptic strength of the afferents of biological neurons. Furthermore, the numerical output or activation of artificial neurons represents the action potential of biological neurons. Similar to action potentials that are generated only after the neuron's membrane potential crosses the threshold potential, the output of artificial neurons mimics an active state only after the weighted sum of the inputs crosses an activation threshold (usually 0).

Several activation functions with different properties, advantages, and disadvantages have been proposed over the years. Early ANNs mostly used smooth approximations to on/off switches, such as the sigmoid function, which is bounded between 0 (inhibition) and 1 (activation), and the hyperbolic tangent function, which is bounded between −1 (inhibition) and 1 (activation). The main problem with these functions is that after reaching values close to the active state (~1), their growth rate (gradient) stunts (e.g., a sigmoid would output 0.9933, 0.9954, 0.9999 for inputs 5, 10, and 15, respectively), which stalls ANN training based on gradient descent (see Section 2.2), especially for large networks. The rectified linear unit (ReLU) function was proposed as an alternative (Glorot et al., 2011), which outputs 0 for negative inputs and replicates the input value (identity function) for positive inputs, allowing faster computation due to its simplicity, and better convergence during training due to its non-stunted growth rate for positive inputs. These benefits make ReLU the most popular choice in modern ANNs except in neurons where a bounded output is required for a specific operation (e.g., in gated memory manipulations) or for convergence (e.g., in recurrent loops and adversarial training). Despite the benefits of ReLU, neurons using this activation have the risk of becoming permanently inactive (0) during training, behaving effectively as “dead,” limiting the capacity of ANNs. Several alternatives to ReLU have been proposed (Nwankpa et al., 2018), including adaptive activation functions with parameters that can be adjusted during training (Apicella et al., 2021), however, they have been used to a lesser extent.

Artificial neurons are organized in a series of interconnected layers; the number of intermediate or *hidden* layers between the first and the last determines the *depth* of an ANN. While three or more hidden layers are expected for a DNN, state-of-the-art DNNs may have hundreds of layers organized in specialized modules. During training, the series of layers learn a processing hierarchy, in which the level of abstraction or complexity of the features extracted by the layers increases gradually from the first to the last layers, in a way that loosely resembles the functional hierarchy of the sensory cortices. Using the visual system as an example, the primary visual cortex processes basic-level features such as vertical edges while the fusiform gyrus, a high-level visual association

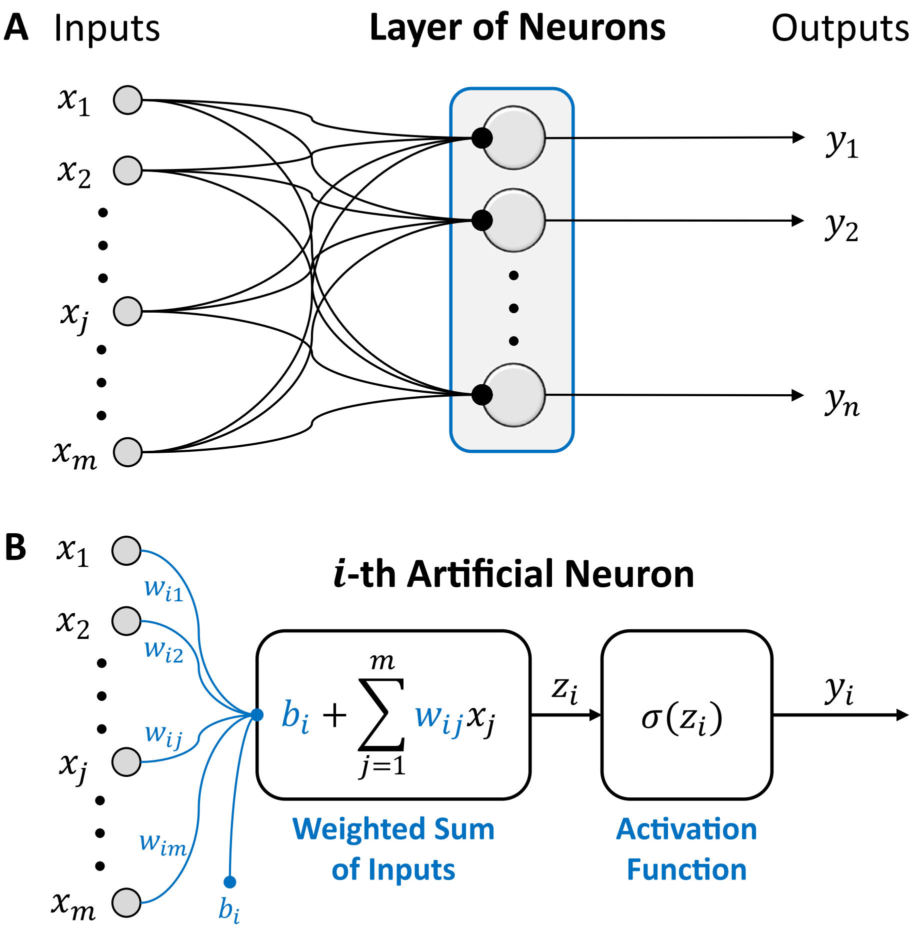


Fig. 1. Basic elements of artificial neural networks. (A) Artificial neural networks are sets of interconnected artificial neurons organized in layers. Each layer receives a multidimensional input and generates a multimodal output with as many dimensions as neurons are in the layer. The multidimensional output is the input of the next layer. (B) An artificial neuron computes a linear combination of the inputs, where the bias and weights are its learnable parameters; the resulting value is fed into a nonlinearity called activation function that generates the neuron's activation (output), loosely resembling the activation of biological neurons.

area, processes faces, a complex type of stimulus.

The processing hierarchy of layers makes DNNs a powerful technique for integrating and extracting information from multimodal data (Fig. 2), i.e., data collected through multiple complementary modalities (e.g., behavioral measures, EEG, and fMRI data). While early layers extract basic-level, modality-specific, "perceptual-like" features, layers located further up the hierarchy extract high-level, abstract or "concept-like" features. Thus, in the case of multimodal data, while each modality (e.g., images, audio, and text) will require a dedicated set of early "perceptual-like" layers (a modality-specific module), high-level layers extracting abstract features can be shared between modalities (modality-free module). Information from each modality would enrich the

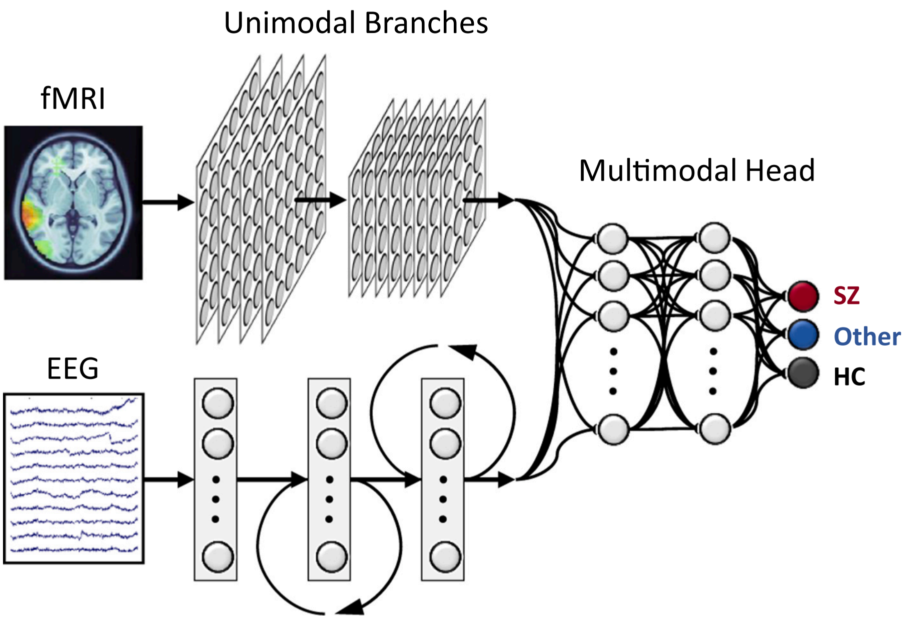


Fig. 2. Deep learning multimodal data fusion. Each data modality requires a dedicated set of layers to extract modality-specific features. The architecture of each unimodal branch is specially designed to suit the characteristics of the data type. In this case, convolutional layers are better suited for fMRI images and recurrent layers are better suited for EEG time series. After the branches, a shared set of layers combines modality-specific features and predicts the class.

abstract concept-like representations of high-level layers, making the DNN more robust in cases of unimodal partial or distorted information. For example, while different sets of perceptual layers are necessary to process the visual and auditory information associated with the shape and sound of a flute, both streams of information can contribute to enriching the concept of a flute.

Based on the direction of the information flow, DNNs can be broadly classified as feedforward or recurrent. In feedforward DNNs, the direction of the information flow is from input to output without feedback loops (e.g., analogous to feedforward connections from V1 to V2 in the visual system). Feedforward multilayered DNNs are universal approximators, i.e., they can approximate any mapping (function) between inputs (e.g., images) and outputs (e.g., categories in a classification task) of a static system with arbitrary precision (Hornik et al., 1989). This type of DNN is most widely used in tasks that do not involve temporal changes (e.g., image recognition). In turn, recurrent DNNs include feedback loops in which layers send feedback information to themselves and/or to layers located earlier in the hierarchy (e.g., analogous to feedback connections from V2 to V1 in the visual system). Analogously to feedforward DNNs, recurrent DNNs are universal approximators of dynamical systems (Schäfer and Zimmermann, 2007). This type of DNN is generally used in tasks involving time-changing or ordinal data (e.g., weather forecasting or language translation). Empirical results have shown that the capacity of DNNs to approximate complex, multivariate, nonlinear systems by far surpasses the results that have been obtained with traditional shallow networks and ML approaches.

One of the reasons for the enormous success of DNNs in solving complex tasks is that, unlike traditional statistical and ML approaches, DNNs are end-to-end approaches, i.e., they not only learn to solve a task (e.g., speech recognition) but also to automatically extract an optimal set of features from the raw data that will be used to solve the task. By learning to extract features directly from raw data, DNNs can overcome some of the limitations and biases affecting manually designed features, resulting in higher performance with less task-specific customization. For example, a DNN architecture that classifies animal species using raw images as input can be trained (without adjustments other than the output/classification layer) to solve a wide range of other tasks such as face recognition, cell type classification, or MRI-based disease diagnosis. The advantages of automatic extraction come at the cost of large training datasets because DNNs have to learn a large number of parameters in order to separate relevant from irrelevant information in the typically high-dimensional and noisy input space.

However, notwithstanding the recent advances in automatic DNN design (Elsken et al., 2018; Zela et al., 2018), three crucial aspects continue to rely largely on human decision: 1) **network architecture**: specific arrangement of neurons and connections determining the flow of information (see Box 1); 2) **learning rules (training algorithm)**: procedures for updating the network weights during training; and 3) **objective functions**: measures of performance or cost associated with an output (e.g., error, reward) that DNNs learn to minimize or maximize during training (Richards et al., 2019). These aspects are designed to address specific characteristics of the task at hand. For example, convolutional layers are usually included in the design of DNNs for image recognition tasks because they are tailored to extract features that are invariant to translation, which is especially suitable for extracting visual elements that remain the same irrespective of their location in an image. Thus, a convolutional layer located high up in the layer hierarchy of a DNN trained for tumor detection would be able to locate a tumor in a CT scan irrespective of the tumor's location in the image.

2.2. DNN training

Error backpropagation (BP) (Rumelhart et al., 1986) is the most widely used algorithm for training DNNs (Fig. 3). In its simplest form, BP can be characterized as a two-step procedure: In the first or **forward step**, the network is fed an input (a training example), predicts an output, the output is compared to the correct output (e.g., a label), and a prediction error (a measure of the difference between the correct and predicted output) is calculated. In the second or **backward step**, the contribution of each weight in the network to the prediction error is estimated through the gradient (derivative of the prediction error with respect to each weight), and the prediction error is reduced by adjusting the values of the weights in a direction opposite to the gradient (a process known as "gradient descent"). Nowadays, DNNs are trained using a number of sophisticated variants of this method (e.g., Adam (Kingma and Ba, 2014)).

The capacity of DNNs to learn arbitrarily complex mappings between inputs and outputs, and decision boundaries between categories, is a double edge sword. While it allows DNNs to reach unprecedented levels of accuracy solving complex tasks, it also makes them vulnerable to poor generalizability caused by overfitting the training dataset, i.e., memorizing the correct answers for a task on the training dataset instead of extracting general relationships or decision boundaries that could be used to solve the same task on data collected independently. To reduce the chances of overfitting, models are usually developed using three datasets commonly referred as training, validation, and test datasets. The **training dataset** is a set of examples used to minimize the model's prediction error for a given task by adjusting the model's *parameters* (e.g., weights of the connections between neurons) using BP or a related training algorithm (Ma et al., 2020). The **validation dataset** is an independent dataset used to find the best performing model by tuning the model's *hyperparameters*, i.e., variables related to the architecture of the network (e.g., adding extra layers), training algorithm (e.g., learning rate in BP), and objective function (e.g., formula to calculate prediction error). Finally, the **testing dataset** consists of data that has not been used for training, adjusting, or validating the model in any way and is used to obtain an unbiased assessment of the model's performance; this assessment can provide information on the model's generalizability.

3. Deep learning research in schizophrenia

In this section, we review deep learning applications for diagnosis (classification) and outcome prediction in the study of psychosis. Our main goal is to explore the potential of DNN algorithms for delivering high-performance models, and thus we focus on articles that used these models rather than traditional ML techniques to solve SZ-related problems. Nevertheless, we include some original applications of traditional ML techniques that could motivate future research.

To make it easier for the reader to get a broad idea of the potential generalizability of the results reported in the text, we classified the model testing procedures of the studies into 2 broad categories (**in-distribution** and **out-of-distribution** testing) with 2 subcategories (**independent** and **non-independent** testing) each (Teney et al., 2020). These categories are based on the statistical relationship between the testing (performance assessment) and both the training (parameter adjustment) and validation (model selection) datasets. Furthermore, we provide information about the sample sizes and cross-validation (CV) schemes used in the studies, if applicable. A detailed list of all the studies on diagnosis and outcome prediction in SZ that were reviewed for this manuscript can be found in Tables 1 and 2, respectively.

The categories for classifying testing procedures were defined in the following way: **In-distribution testing (IDT)** refers to the process of assessing a model's performance on a testing dataset obtained from sampling the same pool of data that was used for building the training and validation datasets; thus, the testing dataset will have the same distribution as the validation and training datasets. The easiest way to conduct IDT on an **independent (I)** dataset (**IDT—I**) is by selecting the cases for training, validation, and testing using a random, non-overlapping partition of all the available data. Unfortunately, it is common "bad practice" to conduct IDT on **non-independent (NI)** datasets (**IDT-NI**) by using testing data for both testing and model selection (validation) purposes. Performance reports obtained with IDT-I

Box 1
Most popular neural network architectures.

Depending on the types of layers used, most neural networks can be broadly classified as multilayer perceptrons (MLPs), convolutional neural networks (CNNs), or recurrent neural networks (RNNs). MLPs are feedforward neural networks consisting of fully-connected layers of artificial neurons (perceptrons). The term fully-connected indicates that all neurons in a given layer are connected to all neurons in the immediately upper layer. On the other hand, the architecture of CNNs is inspired in the visual system of mammals, following the work of Hubel and Wiesel (1968). It has been tailored to include invariances to translation, scaling, and distortion. CNNs are composed of layers of convolutional filters that focus on receptive fields, similarly to biological neurons in the retina and visual cortices. The convolutional layers extract a hierarchy of features. To solve a prediction or classification task in an end-to-end manner, a few layers of a fully-connected ANN (an MLP) or recurrent ANN are typically added on top of the convolutional layers. Finally, RNNs have recurrent connections that allow ANNs to use past outputs in addition to current inputs. Modern RNNs, such as Long Short-Term Memory (LSTM) ANNs and Gated Recurrent Unit (GRU) ANNs, include memory cells that are protected from irrelevant perturbations through gates. Usually, these gates regulate the read and write access to a memory cell when a new input arrives and allow a memory cell to be reset when its content becomes obsolete. In this way, modern RNNs can learn long term dependencies more easily.

Two or more ANNs of any of the three types mentioned previously can be combined into a single model with specialized subnetworks. Popular examples are generative adversarial networks (GANs) and autoencoders. GANs consist of two ANNs, the generator and the discriminator, that compete in a zero-sum game. The generator ANN generates data from noise, and the discriminator ANN takes as input both generated data and true data. In this game, the discriminator tries to differentiate between the training real data and the generated data, and the generator tries to fool the discriminator. After successful training, the generator of a GAN can produce synthetic data with a distribution that is very similar to the training dataset. On the other hand, autoencoders consist of two neural networks: an encoder and a decoder. The encoder maps the input to a latent representation of lower dimensionality than the input, and the decoder takes the latent representation and outputs an approximation of the original input. This is an example of unsupervised learning, where no class labels are required. After training, the encoder can be used to obtain a compact representation of a dataset, useful for dimensionality reduction. A more sophisticated encoder-decoder architecture is the variational autoencoder (VAE), where the encoder maps the inputs into a distribution rather than a single point, from which latent representations can be sampled.

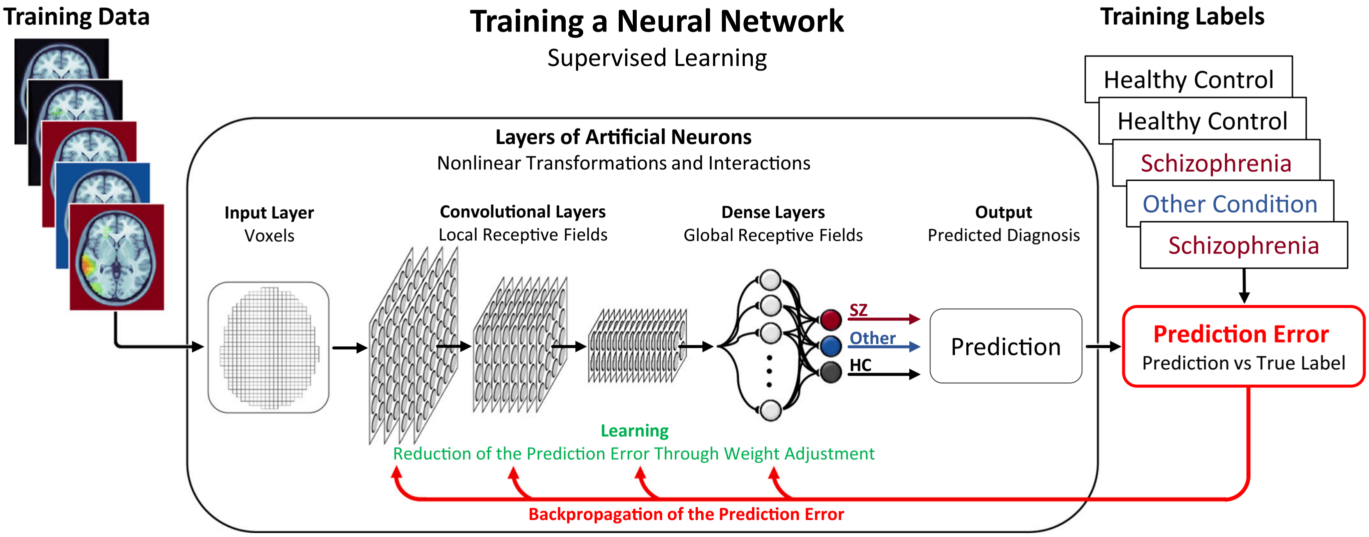


Fig. 3. Deep neural network for supervised learning.
The figure shows a typical deep neural network for image-based diagnosis (classification) with several layers of artificial neurons organized in a processing hierarchy. First, convolutional layers composed of neurons that interact with small input regions (local receptive fields) automatically extract optimal features from raw data. Next, dense or fully-connected layers composed of neurons that interact with the entire input (global receptive fields) differentiate between the available class categories. In supervised learning, a training dataset of examples with their corresponding classes (labels) is required, and the connections (weights) of the neural network are adjusted by minimizing the discrepancy between the prediction and the label (prediction error) using error backpropagation.

are more likely to reflect the real performance of a model on new data (generalization) than reports obtained with IDT-NI. **Out-of-distribution testing (ODT)** refers to the process of assessing model performance using testing datasets collected under different conditions (e.g., at another institution, with equipment from a different manufacturer, etc.) than the training and validation datasets. While measurements continue to be the same across the training, validation, and testing datasets (e.g., resting state EEG), changes in the data collection conditions can shift the distribution of the testing data in ways that have been shown to render state-of-the-art diagnostic DL models unusable, when deployed in real clinical settings and fed data collected in situ (Beede et al., 2020; Taori et al., 2020). Thus, depending on the characteristics of the data, ODT conducted on **independent** (**ODT—I**) datasets can provide one of the best assessments of a model's generalizability. Unfortunately, it is common among studies using ODT to select models (validation) using all the data available, including the testing dataset. This bad practice leads to conducting ODT on **non-independent** (**ODT-NI**) datasets and, therefore, reduces the chances of getting an accurate estimate of a model's generalizability. In the following sections, when a study did not provide enough information to determine the type of testing that was

Table 1
Reviewed studies focused on SZ diagnosis.

Reference	Classification task	Data origin	Sample (subjects)	Task protocol	Input features	ML algorithm	Performance
Functional MRI and genomics data							
Li et al. (2020)	SZ vs HC.	1 site.	183	SNP genotyped from blood sample. fMRI during sensorimotor task (block-design motor response to auditory stimulation).	Loci of quality-controlled SNP data, and ROIs of fMRI voxels given by the AAL template.	Canonical Correlation Analysis on features learned by two fully-connected, sparse autoencoders (one for each domain), followed by SVM.	95.7 ± 0.1% accuracy for SNP domain; 80.5 ± 0.2% accuracy for fMRI domain.
Functional MRI and structural MRI data							
Salvador et al. (2019)	SZ vs HC.	1 site.	211	fMRI from both resting-state and during working memory task (n-back task).	Five 2D brain maps: gray matter voxel-based morphometry from sMRI; regression coefficients for 1back-vs-baseline and 2back-vs-baseline from task-based fMRI; amplitude of low-frequency fluctuations and weighted global brain connectivity maps from resting-state fMRI.	1D convolutional ANN applied to the concatenated brain maps.	84% accuracy.
Functional MRI data							
Kim et al. (2016)	SZ vs HC.	1 site.	100	Resting state and eyes open.	Functional connectivity matrices computed from correlation of group ICA time courses, as 1D vector.	Fully-connected ANN initialized from fully-connected, sparse autoencoder with adaptive sparsity level.	86.5 ± 1.2% accuracy.
Patel et al. (2016)	SZ vs HC.	1 site.	80	Resting state and eyes open staring at a fixation cross.	ROIs of active gray matter voxels according to covariance analysis and AAL template.	SVM on the concatenated features learned by several fully-connected, sparse autoencoders (one for each ROI).	92% accuracy.
Dakka et al. (2017)	SZ vs HC.	1 site.	95	Auditory oddball task.	Voxels of 4D fMRI.	Recurrent ANN with LSTM layers.	66.4% accuracy.
Zeng et al. (2018)	SZ vs HC.	7 sites.	734	1st to 6th site: resting state and eyes closed. 7th site: three working memory tasks.	Three functional connectivity matrices (multi-atlas), as 1D vectors.	SVM on features learned by a fully-connected, sparse autoencoder with inter-class correlation penalization. Three such models, one for each input type, and decision by majority vote.	85 ± 1.2% accuracy when pooling sites; 81.0 ± 4.9% leave-one-site-out accuracy.
Wang et al. (2019)	SZ vs HC.	1 site.	131	Resting state and eyes open staring at a fixation cross.	Functional connectivity matrices based on AAL template.	Capsule ANN.	82.4% accuracy.
Niu et al. (2019)	SZ vs HC.	1 site.	82	Resting state.	fMRI frame with augmentation techniques: different ICA orders to extract Default Mode Network component, and different spatial smoothing parameters after ICA.	2D convolutional ANN.	90.8% accuracy.
Yang et al. (2019)	SZ vs HC.	3 sites.	222	Resting state.	Three input types: coefficients from learned dictionaries with different sparsity regularizations; distances to other samples using Gaussian kernels with different parameters; functional connectivity matrices based on AAL template.	Three capsule ANNs (one for each input type), and decision by a weighted average of outputs.	82.8 ± 7.6% accuracy when pooling sites.
Lei et al. (2019)	SZ vs HC.	5 sites.	747	Resting state.	Functional connectivity matrices based on AAL template, as 1D vector.	Fully-connected ANN.	81 ± 2% accuracy when sites are isolated; 62.5–68.1% leave-one-site-out accuracy.
Matsubara et al. (2019)	SZ vs HC; BD vs HC.	1 site.	165 for SZ; 63 for BD.	Resting state.	Single fMRI frame, using average voxel value of ROIs given by the AAL template.	Fully-connected conditional variational autoencoder.	SZ: 71.3% balanced accuracy. BD: 64%

(continued on next page)

Table 1 (*continued*)

Reference	Classification task	Data origin	Sample (subjects)	Task protocol	Input features	ML algorithm	Performance
							balanced accuracy.
Oh et al. (2019)	SSD vs HC.	1 site.	144	Audiovisual stimuli task (evoked negative and neutral emotion).	3D activation map constructed based on the 4D fMRI data.	3D convolutional ANN initialized from pretrained convolutional autoencoder.	84.2% accuracy.
Qureshi et al. (2019)	SZ vs HC.	1 site.	144	Resting state and eyes open staring at a fixation cross.	3D volumetric images from group ICA decomposition.	3D convolutional ANN.	98 ± 1% accuracy.
Yan et al. (2019)	SZ vs HC.	7 sites.	1100	Resting state.	Time courses of ICs.	Convolutional and recurrent ANN: 1D convolutional layers followed by GRU layers.	83.2 ± 3.2% accuracy when pooling sites; 80.2 ± 3.0% leave-one-site-out accuracy.
Zhao et al. (2020)	SZ vs HC; MDD vs HC.	7 sites for SZ; 4 sites for MDD.	1100 for SZ; 555 for MDD.	Resting state and eyes closed.	Functional connectivity matrices computed by correlation of ICs, as 1D vector.	Fully-connected generative adversarial network (GAN).	SZ vs HC: 82.1 ± 0.7% accuracy when pooling sites; 80.7 ± 3.8% leave-one-site-out accuracy. MDD vs HC: 70.1 ± 0.6% accuracy when pooling sites; 64.3 ± 2.9% leave-one-site-out accuracy.
Structural MRI data							
Pinaya et al. (2016)	SZ vs HC; tested on FEP.	1 site.	226; And 32 FEP.	–	Cortical thickness of brain regions and volumes of anatomical structures from Desikan-Killiany atlas.	Fully-connected ANN initialized from a deep belief network.	73.6 ± 6.8% balanced accuracy. Only 56.3 ± 6.8% held-out FEP subjects predicted as SZ.

Pinaya et al. (2019)	SZ vs HC;
	ASD vs HC.
		1 site for classification;
		1 site for HC pretraining.
			75 for SZ;
			188 for ASD;
			1113 for HC pretraining.
				–	Cortical thickness of brain regions from Desikan-Killiany atlas, and volume of neuroanatomical structures from whole-brain segmentation.	Fully-connected autoencoder that also encodes age and sex, pretrained on HC data. Prediction based on reconstruction error.	SZ: 0.707 AUC.
							ASD: 0.639 AUC.

Vieira et al. (2020)	FEP vs HC.	5 sites.	956	–	Volume and thickness of predefined cortical and subcortical regions extracted with FreeSurfer.	PCA followed by fully-connected ANN.	63 ± 6% balanced accuracy when sites are isolated. Poor cross-site performance.
Oh et al. (2020)	SZ vs HC.	6 sites.	926	–	Voxels from 3D images.	3D convolutional ANN.	88.6% accuracy when pooling sites; 70% accuracy when evaluating on held-out site.
Genomics data							
Chen et al. (2018)	SZ vs HC.	3 sites.	13,585	Genomic data (SNPs) from three SZ case-control studies. Genetic markers associated with SZ and 28 comorbidities were identified from the literature.	Independent polygenic risk scores for SZ and each comorbidity.	Fully-connected ANN.	72.1% accuracy.
Wang et al. (2018)	SZ vs HC.	1 site.	710	Genomic (SNPs) and transcriptomic (gene expression, enhancer H3K27ac activation levels, cell fraction estimates, and co-expression module mean expression) data from Prefrontal Cortex.	Binarized genomic and transcriptomic data by thresholding at the median value.	Fully-connected, conditional deep belief network with sparse connectivity restrictions based on genomic analysis.	73.6% accuracy.

(continued on next page)

Table 1 (*continued*)

Reference	Classification task	Data origin	Sample (subjects)	Task protocol	Input features	ML algorithm	Performance
EEG data							
Calhas et al. (2020)	SZ vs HC.	1 site.	84	Resting state and eyes closed.	Time-frequency images using STFT.	2D convolutional siamese ANN followed by XGBoost.	95 ± 5% accuracy.
Phang et al. (2020)	SZ vs HC.	1 site.	84	Resting state and eyes closed.	Three input types: time-domain VAR model coefficients matrix (2D); frequency-domain PDC matrix (2D); hand-crafted complex network measures (1D).	Three convolutional ANNs (one for each input type), and decision by a weighted average of outputs.	91.7% accuracy.
Interview audio data							
Naderi et al. (2019)	SZ vs MDD vs BD vs HC.	1 site.	363	Audio recording when talking about their children without interruption.	Audio and text transcription, divided in multiple segments based on changes in speech.	Several recurrent and convolutional ANNs to extract generic language features and emotion-specific features from audio and text independently. The features are concatenated and processed by a recurrent ANN.	74.4% average accuracy.

Table 2
Reviewed studies focused on SZ prediction.

Reference	Task	Sample	Task protocol	Input features	ML algorithm	Performance
Electronic health records (EHR) data						
Miotto et al. (2016)	One-year new disease diagnostic prediction over 78 diseases, including SZ.	794,587 subjects for ANN training; 281,214 subjects for ML classifier.	EHR from Mount Sinai data warehouse.	Frequency of structured fields, and multinomials over 300 automatically extracted topics in clinical narratives.	Random forest classifier on features learned by a fully-connected autoencoder.	92.9% average accuracy.
Holderness et al. (2019)	Detection of readmission risk factor domains in clinical notes of FEP patients.	2,100,000 sentences for training (RPDR); 4847 sentences for testing (McLean).	EHR from Research Patient Data Registry (RPDR) and McLean Meditech.	Words of a sentence.	Fully-connected ANN on the average USE word embedding of the sentence.	82.8% macro-average F1-score.
Senior et al. (2020)	Detection of OxMIS's suicidal risk factors in clinical notes of SZ and BD patients.	308 documents containing 10,151 annotated text spans.	EHR from Oxford Health NHS Foundation Trust.	Words of a sentence.	Convolutional ANN on the GloVe word embeddings of the sentence.	83% micro-average F1-score.
Interview transcription data						
Rezaii et al. (2019)	Prediction of conversion to psychosis in prodromal subjects.	40 subjects; 30,000 Reddit users for feature extraction.	Transcription of audio recorded during Structured Interview for Prodromal Syndromes (SIPS). Subjects were followed up for 2 years or to conversion to SZ.	Words of a sentence.	Logistic regression on two features extracted from the word2vec word embeddings of the sentence.	90% accuracy.
Mental health journal data						
Shickel et al. (2017)	Sentiment classification of responses sent to an online therapy service.	1.6 million tweets for training; 3872 patients' responses for testing.	Patients' responses consist of daily thoughts and feelings.	Words of a document.	Recurrent ANN with GRU layers and attention mechanism, on GloVe word embeddings of the document.	78% accuracy.
EEG data						
Ahmedt Aristizabal et al. (2020)	Detection of children at risk of conversion to SZ (RSZ).	105 subjects.	Auditory oddball task. RSZ children underwent three assessments: an initial one (A1), a 2-year follow-up one (A2), and a 4-year follow-up one (A3).	Raw EEG signals.	Convolutional and recurrent ANN: 2D convolutional layers followed by LSTM layers.	When trained in A1: 72.5% accuracy in A1, 69.8% accuracy in A2, 67.0% accuracy in A3.
Fernando et al. (2020)	Detection of children at risk of conversion to SZ (RSZ).	104 subjects.	Auditory oddball task.	Raw EEG signals.	Recurrent ANN: LSTM layers followed by a Neural Memory Network with plasticity mechanism.	93.9 ± 0.2% accuracy.

used, we state the possible alternatives (e.g., "IDT-NI or I").

3.1. Patient diagnosis – pattern recognition and classification

In general, most deep learning studies in SZ have been focused on solving the binary classification task (pattern recognition) of differentiating between SZ patients and healthy controls (Fig. 4). These studies utilized, in decreasing order of popularity, fMRI, MRI, genomics, and EEG data (Table 1). While one of the strengths of DNNs is their ability to automatically extract useful features directly from raw data, with the exception of an early study using convolutional and recurrent DNNs on fMRI data with poor results (66.4% accuracy, IDT-NI, n = 95, 10-fold CV) (Dakka et al., 2017), most studies using DNNs for SZ diagnosis have used manually-designed features. For instance, fully-connected ANNs, a basic type of ANN with limited feature extraction capabilities, have been used, with varying levels of success, to differentiate between SZ and healthy controls using features such as MRI-extracted cortical thickness and structural volume (three studies: 73.6% accuracy, IDT-NI, n = 226, 3-fold CV; 0.71 AUC[2] ODT—I, n = 75; 63% accuracy, IDT—I, n = 191, nested 10-fold CV) (Pinaya et al., 2016; Pinaya et al., 2019; Vieira et al., 2020), fMRI-extracted functional connectivity (six studies: 81 to 86.5% accuracy, IDT-NI, n = 100 to 1100, 5-fold or 10-fold CV) (Kim et al., 2016; Lei et al., 2019; Wang et al., 2019; Yang et al., 2019; Zeng et al., 2018; Zhao et al., 2020), and fMRI-extracted, atlas-based mean ROI activations (two studies: 71.3 to 92% accuracy, IDT-NI, n = 60 to 165, 10-fold CV) (Matsubara et al., 2019; Patel et al., 2016).

During the last couple of years, SZ research, like many other areas of research, has seen a rapidly increasing number of studies using deep learning methods on minimally-processed data. These studies have replicated the gains in prediction and classification accuracy resulting from using DNNs on minimally-processed data, that have been observed in other areas of research (e.g., computer vision). By replacing hand-engineered (e.g., theory-oriented) features with data-driven ones, deep learning may provide a way to uncover crucial patterns in the data that have been hiding in plain sight due to pre-existing assumptions and expectations about the disorder.

DNNs have been used for SZ diagnosis (classification) on neuroimaging data with very good results (accuracy ~90%). For example, 2D convolutional DNNs have been used on fMRI single-frame data (90.8% accuracy, IDT-NI, n = 82, 5-fold CV) (Niu et al., 2019) and EEG time-frequency (95% accuracy, IDT-NI, n = 84, leave-one-out CV) (Calhas et al., 2020) and connectivity (91.7% accuracy, IDT-NI, n = 84, 5-fold CV) (Phang et al., 2020) data. Accurate results have also been obtained using 3D convolutional DNNs on MRI voxel data (88.6% accuracy, IDT—I, n = 866, 10-fold CV) (Oh et al., 2020) and fMRI 3D map (84.2% accuracy, IDT-NI, n = 144, 10-fold CV) (Oh et al., 2019) and volumetric (98% accuracy, IDT-NI, n = 144, 10-fold CV) (Qureshi et al., 2019) data. Finally, recurrent DNNs have obtained high accuracy harnessing the brain dynamics captured in fMRI time series data (83.2% accuracy, IDT-NI, n = 1100, 5-fold CV) (Yan et al., 2019).

Just a handful of studies have used deep learning on data modalities other than neuroimaging. Studies using genetic data, such as polygenic risk score for SZ and comorbidities (Chen et al., 2018) and prefrontal genomics and transcriptomics (Wang et al., 2018), have shown that deep learning matched (72.1% accuracy, ODT—I, n = 1492) (Chen et al., 2018) or surpassed (73.6% accuracy, IDT-NI, n = 710, 10-fold CV) (Wang et al., 2018) linear models' performance. In a recent and highly novel study, Naderi et al. (2019) reached relatively high accuracy (74.4%, IDT-NI or I, n = 363, 5-fold CV) using convolutional and recurrent DNNs in combination with a random forest algorithm to diagnose mental disorders (SZ, major depressive disorder, and bipolar disorder) from audio recordings of speech. Importantly, this study took advantage of large pretrained models for audio and text processing that are openly available online. The easy and open access to pretrained models has an enormous potential for developing novel diagnostic tools, such as language-based systems based on large, state-of-the-art natural language processing (NLP) models (Brown et al., 2020b; Devlin et al., 2018).

As mentioned, deep learning is an extremely powerful tool for extracting, integrating, and using information from multimodal data (e.g., simultaneous EEG/fMRI); however, few studies have harnessed this capability in SZ research. For example, Salvador et al. (2019) used a small 1D convolutional ANN on brain MRI/fMRI maps, achieving accuracies comparable to traditional ML methods (84% accuracy, IDT-NI, n = 211, 10-fold CV). Also, Li et al. (2020) used a multi-step approach in which fMRI and genomic data were fed to fully-connected ANNs followed by canonical correlation analyses (CCA). Each data modality was decomposed into components (projections) containing high-level features that were common to both modalities. Each modality's components were used independently to classify subjects, achieving competent fMRI-based accuracy (80.5%, IDT-NI, n = 55) and high genomic-based accuracy (95.7%, IDT-NI, n = 55).

3.2. Prediction of psychosis and early diagnosis

SZ is a disabling, chronic mental disorder that usually starts in adolescence, following a prodromal phase with attenuated SZ-like symptomatology (Fusar-Poli et al., 2012). It is estimated that ~30% of patients in a prodromal phase convert to psychosis within a 3-year follow-up period (Fusar-Poli et al., 2012). The course of schizophrenia is punctuated by relapses (Robinson et al., 1999), with 80–90% of patients experiencing one or more clinical decompensations within 5 years of the first episode (Robinson et al., 1999; Zipursky et al., 2014). Developing deep learning systems for monitoring the risks of conversion and relapse could allow the implementation of timely interventions to halt the onset and relapse of psychosis, or to minimize their consequences (Barnes et al., 2008; Marshall et al., 2005).

A nontrivial problem of any forecasting system is the selection of a set of predictors suitable for anticipating an event within a certain timeframe (e.g., lifespan, six months, a few weeks, etc.). Most studies have used language to predict psychosis outcomes, under the assumption that the way in which someone's speech diverges from *normal* speech is an early sign of impending psychosis (first episode or decompensation). For example, Rezaii et al. (2019) used transcriptions of the Structured Interview for Prodromal Syndromes (SIPS) conducted on a small sample of young adults at clinical high risk of psychosis participating in the North American Prodrome Longitudinal Study (NAPLS). Using a logistic regression model, they found that low semantic density (poverty of content) and talk about voices and sounds predicted the onset of psychosis within a 2-year period with high accuracy (90%, IDT—I, n = 10). These predictors were extracted from word *embeddings* learned automatically by an ANN trained on a large corpus of text (Word2Vec) (Mikolov et al., 2013a,b). Word embeddings map the words of a language into a vector space of reduced dimensionality that preserves word associations (i.e., words that occur in similar contexts are represented by similar vectors). The advantage of using word embeddings is that they allow for the use of a rich trove of mathematical vectorial operations in the context of language analysis.

Other studies have used electronic health records (EHRs) with very good results. The advantage of using EHRs is that there are enormous databases available online. Miotto et al. (2016) trained fully-connected ANNs on over one million patients to encode each patient's entire EHR into a vector that they called *Deep Patient*. Using a random forest classifier on Deep Patient vectors, the authors were able to predict the diagnosis of 78 diseases, including SZ, within a 1-year time range with

[2] AUC stands for area under the Receiver Operating Characteristic (ROC) curve. It summarizes the overall performance of a binary classifier on a scale from 0.5 to 1, where 1 would correspond to a perfect classifier. The ROC curve is generated by plotting the sensitivity versus the false positive rate (1 minus specificity) of a classifier for every possible discrimination threshold.

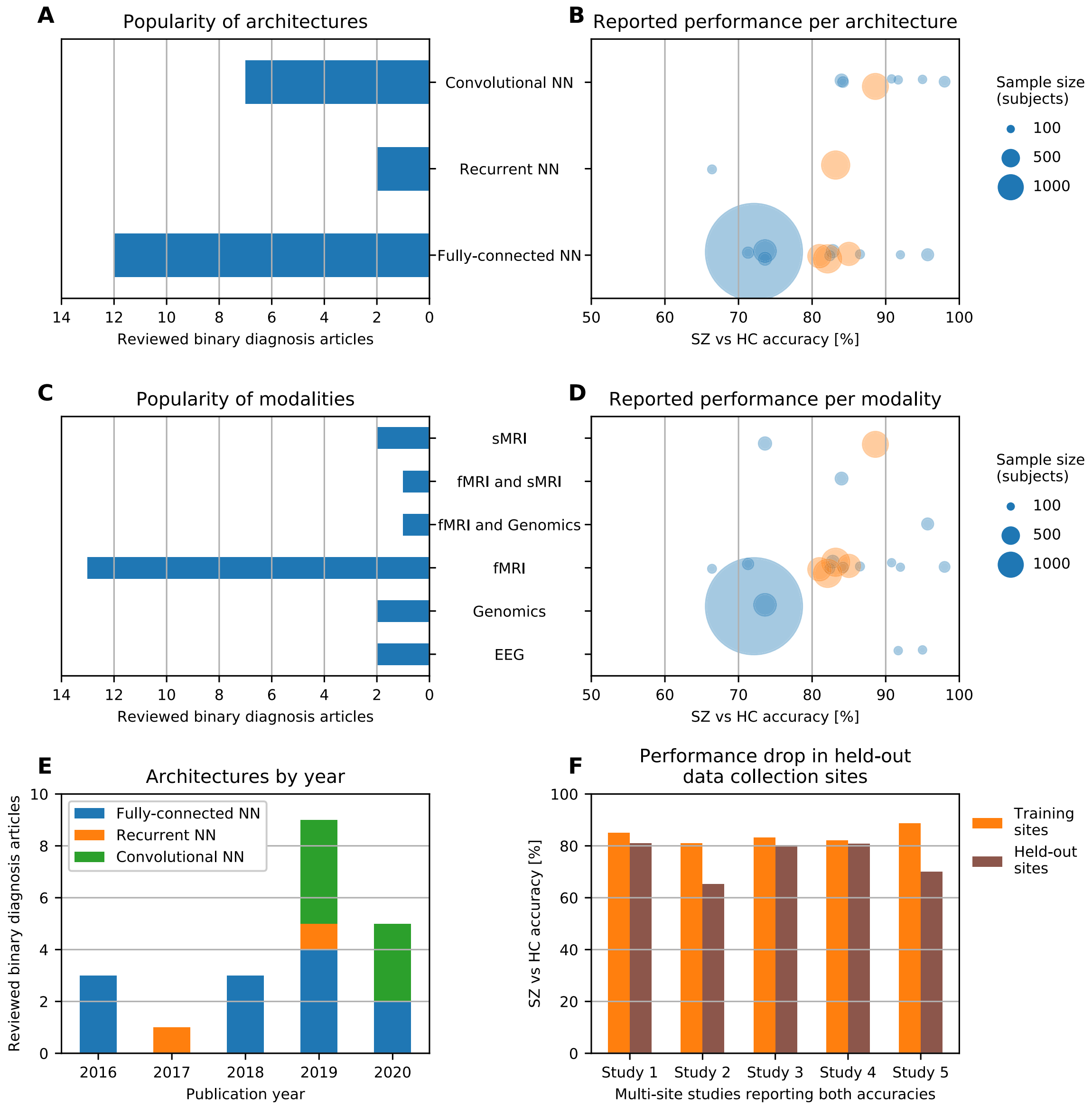


Fig. 4. Review of binary diagnosis of schizophrenic (SZ) patients against healthy controls (HC).
(a) Number of reviewed articles by ANN architecture. (b) Reported accuracy of the binary diagnosis task by ANN architecture. (c) Number of reviewed articles by data modality. (d) Reported accuracy of the binary diagnosis task by data modality. (e) Number of reviewed articles by ANN architecture and publication year. (f) Comparison of the accuracy reported by studies using datasets from multiple data collection sites, when models were evaluated on data from collection sites used during training (pooled sites evaluation) or on data from held-out sites that were not used during training (leave-one-site-out evaluation). In panels (b) and (d), the size of the circles represents the sample sizes (number of subjects) of the studies, and the orange circles highlight the five studies showed on panel (f).

an overall accuracy of 92.9% (IDT—I, $n = 76,214$).

Few studies have used EEG data for prediction tasks (for details, see Table 2). Two studies used auditory mismatch negativity (MMN) EEG data collected longitudinally on children at risk of SZ and controls (Ahmedt Aristizabal et al., 2020; Fernando et al., 2020). MMN is an event-related potential response to oddball stimuli associated with cognitive dysfunction in SZ (Lee et al., 2017). One of the studies used a recurrent convolutional ANN to train and test on data from three time points: baseline, 2-year follow-up, and 4-year follow-up (Ahmedt Aristizabal et al., 2020). After training the model on baseline data alone, it achieved accuracies of 72.5% (IDT-NI, $n = 105$, 5-fold CV), 69.8% (ODT-NI, $n = 110$), and 67.0% (ODT-NI, $n = 99$) on testing data from the baseline, 2-year, and 4-year timepoints, respectively. In the other study, a type of recurrent ANN known as neural memory network was combined with a plasticity mechanism that strengthens or weakens the model's neuronal connections based on experience beyond training (Fernando et al., 2020). When trained and evaluated on data from the baseline period, the model achieved 93.9% accuracy (IDT-NI, $n = 104$, 5-fold CV).

Instead of directly predicting risk of diagnosis, some studies have built systems to extract useful predictors from data to be used by clinicians or in combination with other tools to assess the risk of disease outcomes such as readmission, suicide, etc. For example, one study focused on the problem of readmission after discharge. They trained a fully-connected ANN to map sentences in EHRs from first-episode patients to 7 readmission risk factor domains (Holderness et al., 2019). After training, they obtained good results, achieving 82.8% F1-score[3] (IDT—I, $n = 4847$ sentences), compared to the risk factors identified by

[3] F1-score is a statistical measure of classification performance that balances false positives and false negatives.

clinicians. A study developing tools for assessing suicide risk within a 1-year timeframe fine-tuned a pretrained convolutional ANN to extract 17 relevant variables from EHRs of SZ-spectrum and bipolar disorder patients. After training, the model achieved an 83% F1-score (IDT—I, $n =$ 1055 text spans) compared to variables extracted manually (Kormilitzin et al., 2020; Senior et al., 2020). Finally, a study classified sentiments from online mental health journals for risk assessment and well-being monitoring (Shickel et al., 2017). They fine-tuned a recurrent ANN pretrained on tweets, achieving 78% accuracy (IDT—I, $n = 3872$ responses, 5-fold CV).

4. Insights into the mechanisms of schizophrenia through the lens of deep learning

The so called "group of schizophrenias" is a highly complex disorder at multiple levels. First, evidence suggests that there is not one unique causal path to SZ but many, each one comprising several risk factors interacting with each other. Second, SZ symptoms are heterogeneous among patients and involve almost every aspect of the human mind, including language (e.g., verbal hallucinations and poverty of speech), emotions (e.g., blunted affect), volition (e.g., avolition), cognition (e.g., working memory deficits), and motion (e.g., motor disinhibition), among others. Finally, SZ is a psychiatric disorder that concerns the brain, which is a high-dimensional nonlinear system at multiple levels (e.g., single neurons, local circuits, nuclei such as the striatum, large scale functional networks, etc.). Traditional statistical and ML methods have failed to model and harness the information contained in this complexity. While existing deep learning approaches are still insufficient to completely model systems like this, they offer a novel way to predict and better understand SZ.

4.1. The black box notion of deep neural networks

As described earlier, it is commonly stated that one of the problems of DNNs is that, despite making remarkably accurate predictions, they are black boxes, i.e., the internal mechanisms underlying DNNs' outcomes are either unknown or too complicated to be interpreted in any meaningful way (Lipton, 2018). DNNs usually take thousands of input variables (e.g., a small grayscale image of 128 × 128 pixels has 16,384 pixels or input variables), which are combined and transformed nonlinearly multiple times using thousands (tens of billions in large, state-of-the-art DNNs) (Brown et al., 2020a) of trainable parameters (weights) to generate the outputs. Thus, it is not possible to obtain the kind of transparency in DNNs that linear models have, in which outputs are an interpretable linear combination of inputs. Another factor contributing to the lack of (algorithmic) transparency of DNNs is that, similarly to the nervous system, DNNs process information using distributed representations, i.e., each "concept" is represented by many neurons, and each neuron participates in the representation of many concepts (Roy, 2012). Thus, while it is possible to identify what kind of concept is being represented by some groups of neurons (e.g., neurons that respond exclusively to vertical edges), this is not always the case, especially for neurons or layers higher up in the processing hierarchy of a DNN.

4.2. Extracting insights from deep learning predictions

In fields like healthcare, in which decisions involve the well-being of human beings, clinicians and scientists may be reluctant to entrust decisions to algorithms without having a clear understanding of the criteria or variables involved in making those decisions (Lipton, 2018). There is no easy way to make a DNN transparent without simplifying it in a way that may compromise its performance (e.g., by using less input variables and neurons). However, there are many retrospective (post hoc) methods aimed at explaining a DNN's output *after* the output has been generated (for a review, see (Arrieta et al., 2020)).

After processing an input (e.g., an MRI scan from a patient), post hoc methods allow researchers to determine what variables in the input (e.g., what voxels in an MRI scan) influenced the DNN's output (e.g., patient classified as SZ) the most. While these methods do not provide a rationale, justification, or mechanistic explanation for a DNN's output, by uncovering the most relevant input variables, these methods are equipping clinicians and scientists with the basic building blocks to advance novel explanatory models and testable hypotheses. For instance, imagine a DNN trained to compare pairs of images of fruits and determine whether both fruits are of the same type (Fig. 5). If the DNN determines that a yellow pomelo and a banana are of different types, a post hoc method would probably reveal that pixels located at the fruits' contours were highly relevant for the output. Thus, we could build an explanatory model saying that the shape, and not the color, is the most relevant trait differentiating images of yellow pomelos and bananas. Instead, if the DNN determines that a yellow pomelo and an orange are of different types, a post hoc method would probably reveal that the most relevant pixels were spread over the interior regions of the fruits. Thus, we could hypothesize that the color, and not the shape, is the most important characteristic differentiating images of these types of fruits. Similarly, DNNs trained in SZ classification tasks can be used as powerful "microscopes" to examine the mechanisms of SZ. For example, a DNN trained to classify two psychiatric conditions using multimodal brain activity data (e.g., simultaneous EEG/fMRI) can be interrogated to reveal how relevant each variable (biomarker) is for classifying (differentiating) the conditions. In this case, while the magnitude of an input's relevance in a DNN can be interpreted analogously to the magnitude of a regression coefficient (β) in a linear regression, the relationship between inputs and output represented by the magnitude is completely different in both cases: nonlinear with complex high-dimensional interactions in a DNN and linear with simple low-dimensional interactions in a regression.

Post hoc explanations of DNN predictions can be obtained either by methods that are directly applicable to any *trained* DNN without involving architectural modifications (Fig. 6a, b, and c), or by methods that incorporate special architectural designs into DNNs *before training* (Fig. 6d). Popular methods in the former include feature relevance, which estimates the effect that changing each input variable has on a DNN's output, and activity visualization, which visualizes features at the different levels of a network (Olah et al., 2017). The inclusion of attention modules into a DNN design is a widely used method in the latter that allows users to identify the elements in the input that "caught" a DNN's attention the most when it generated a particular output (e.g., a specific diagnosis) (Arrieta et al., 2020; Bahdanau et al., 2014; K. Xu et al., 2015). Another approach in the second group that is becoming increasingly popular is the implementation of generative models (e.g., deep belief networks), a special kind of DNN that, besides learning the distinctive features of each class (e.g., distinctive activation patterns in the fMRIs of SZ patients), learns to *generate* synthetic examples of the classes (J. Xu et al., 2015). The comparison between synthetic examples of different classes can be used to reveal class-specific features.

Several studies differentiating between SZ and healthy controls have tried to identify the inputs (usually brain regions) that were most relevant for making a diagnosis using post hoc explanatory approaches. While post hoc explanations are slowly starting to provide novel insights into the mechanisms of SZ, some of the most used post hoc explanations in SZ research have significant limitations that may have undermined the quality of the information obtained with them. Many of these limitations could be overcome using alternative post hoc approaches developed in other areas of research, such as image recognition and natural language processing.

A group of studies used an approach that turned diagnostic DNN classifiers into simple linear transformations by focusing on the weights multiplying each layer's inputs and ignoring the nonlinear activation functions (e.g., sigmoid) between layers (see Section 2.1) (Kim et al., 2016; Lei et al., 2019; Zeng et al., 2018). Similar to linear and logistic

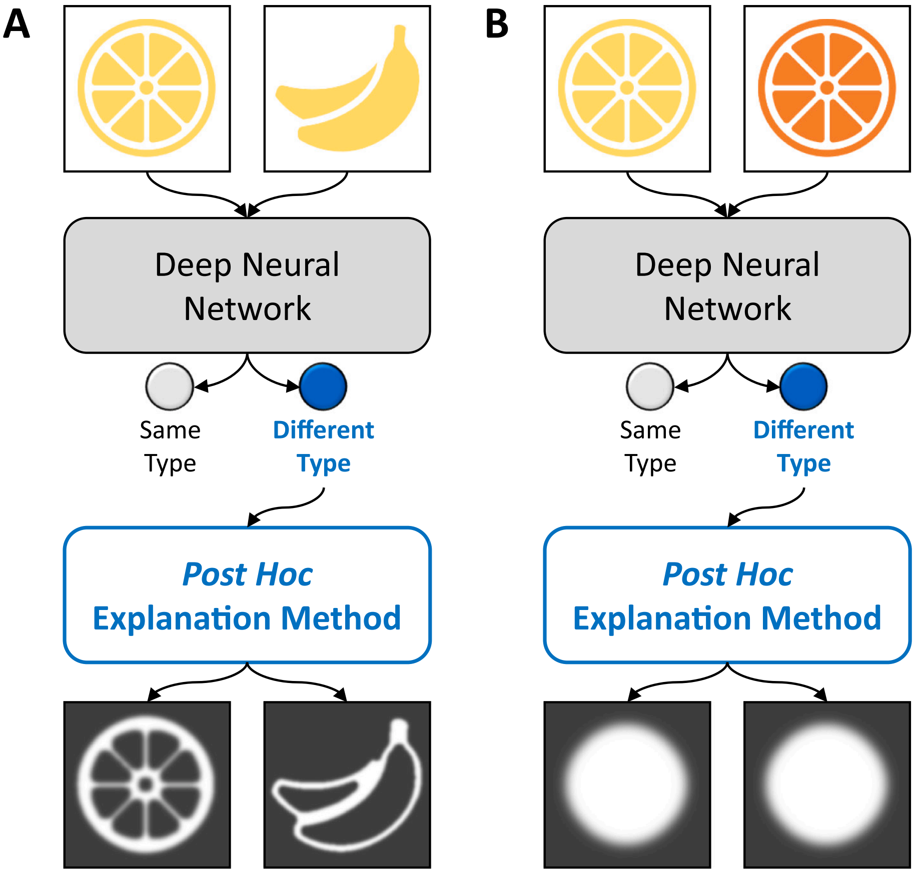


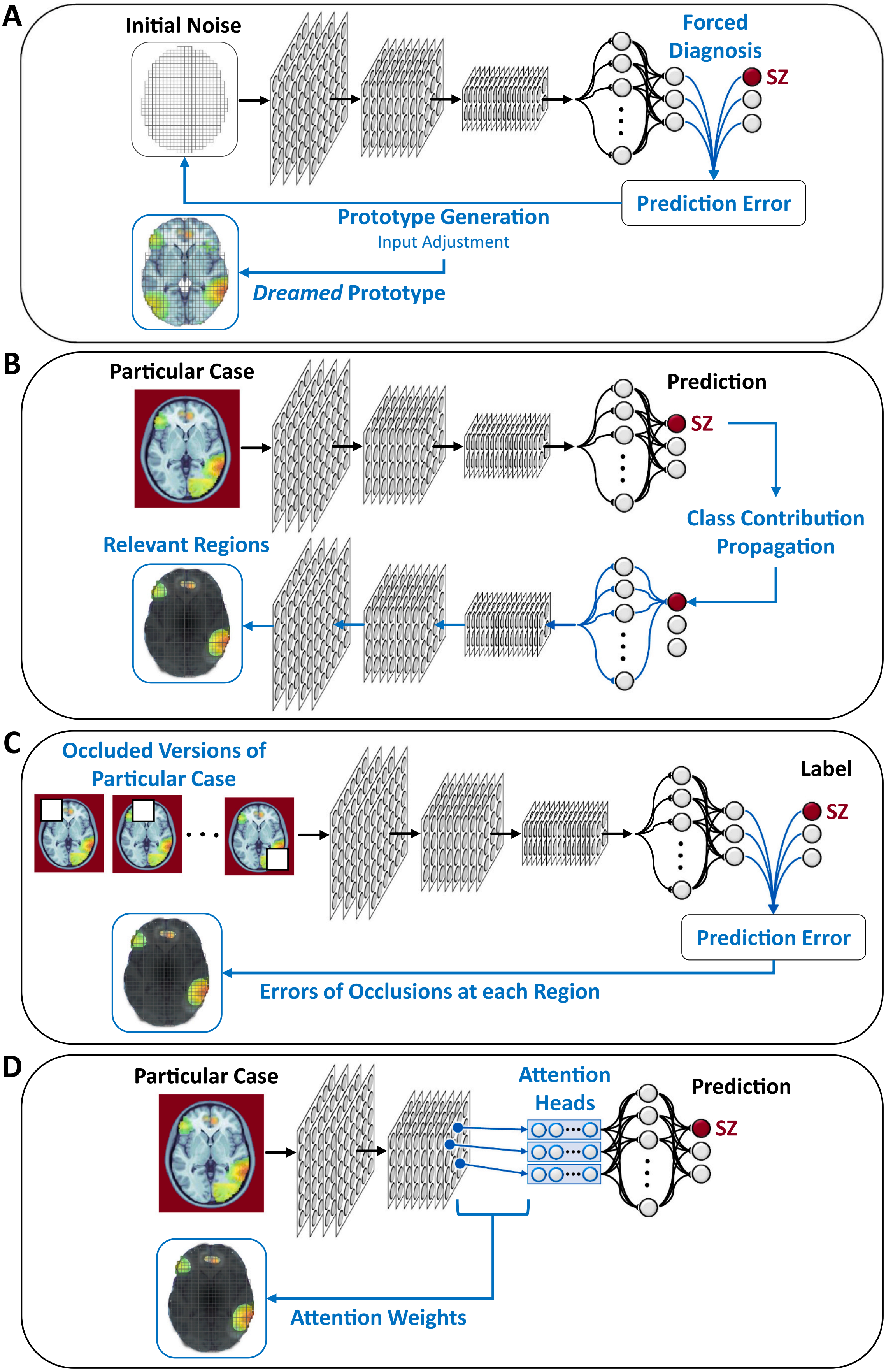
Fig. 5. Hypothesis generation through post hoc explanations.
A deep neural network trained to compare pairs of images of fruits and determine whether both fruits are of the same type. (A) A yellow pomelo and a banana are predicted to be of different types, and a post hoc explanation method applied to this prediction reveals that pixels located at the fruits' contours were highly relevant for the output. Hence, we can hypothesize that the shape, and not the color, is the most important characteristic distinguishing images of yellow pomelos and bananas. (B) A yellow pomelo and an orange are predicted to be of different types as well, but a post hoc explanation reveals that the most relevant pixels were spread over the interior regions of the fruits. Hence, we can hypothesize that this time the color, and not the shape, is the most important characteristic distinguishing images of yellow pomelos and oranges. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

regressions' coefficients (βs), in linearized DNNs the magnitude of the weight multiplying each input (all inputs are in the same scale) is interpreted as the relative relevance of the input (with respect of other inputs) for the network's final output (e.g., diagnosis). Furthermore, like in any linear system, in linearized DNNs the relative relevance (weight magnitude) of an input does not change with (is independent of) changes in the input's and other inputs' values. However, for nonlinear systems like the original DNN, this is not generally true. For example, for a DNN with ReLU activation functions (see Section 2.1), an input's contribution to the output can be effectively zero independently of the magnitude of its weight if the result of the weighted sum of all the inputs is negative (ReLU's output is 0 for negative inputs).

A second group of studies used a method based on the reconstruction error of autoencoders, a type of ANN (see Box 1). In general terms, autoencoders are simultaneously trained to both compress a usually high dimensional input into a comparatively low dimensional encoding and use the encoding to reconstruct the original input with as little error as possible. The studies determined the reconstruction error resulting from feeding neuroimaging data from SZ patients to autoencoders trained or adjusted for processing data from a different diagnostic group (healthy controls) (Matsubara et al., 2019; Pinaya et al., 2019). The magnitude of *each variable*'s (e.g., brain area) reconstruction error was interpreted as a measure of the variable's relevance for differentiating SZ from the other diagnostic group. An important problem with this approach is that it disregards that to compress and reconstruct data, autoencoders may leverage the interdependency between input variables (e.g., correlation between brain areas of the default mode network). This problem may lead to crediting the reconstruction error, and therefore the relevance, to the wrong variables. For example, imagine an autoencoder DNN trained on a dataset of aerial photos taken under normal weather conditions. Suppose the same autoencoder is fed aerial photos of a city that has been flooded by heavy rains in which the roofs of a couple of cars in a parking lot are barely protruding out of the water. Under normal weather conditions, parking lot-sized bodies of water are usually ponds, and no car roofs ever stick out of them. Thus, the autoencoder might encode the image of the flooded parking lot as some sort of pond and, in the process, “clean” out the details that are non-essential for reconstructing a pond (this is the reason why autoencoders are used for image denoising tasks). As a result, the reconstructed images of the flooded parking lot will likely depict a small water surface without car roofs, as expected for a small pond. The comparison between the original and reconstructed images will reveal that the largest reconstruction errors are not located on what is abnormal in the scene (1.5 m of water covering a parking lot), but on the few elements that remain normal (car roofs as seen from above).

A third group of studies used an occlusion approach that assessed the performance drops of a trained DNN classifier resulting from removing (occluding) subsets of input variables (e.g., regions or connections between regions) one at a time. The magnitude of the drop resulting from each occlusion was interpreted as a measure of the potential relevance of the occluded variables in the brain abnormalities underlying SZ (Fig. 6c) (Oh et al., 2020; Phang et al., 2020; Yan et al., 2019; Zhao et al., 2020). However, the occlusion approach is sensitive to the size of the occlusions and, most importantly, completely disregards the effects that interactions or interdependency between inputs may have on determining a diagnostic category. This is analogous to the problems affecting the reconstruction error approach discussed above.

There are a number of alternative approaches from the fields of computer vision and natural language processing that have addressed the aforementioned limitations of the linearization, reconstruction error, and occlusion approaches. We suggest the readers explore techniques such as attention modules (Fig. 6d) (Arrieta et al., 2020; Bahdanau et al., 2014; K. Xu et al., 2015) and generative approaches like the popular DeepDream (Fig. 6a) (Mordvintsev et al., 2015; Simonyan et al.,



(caption on next page)

Fig. 6. Deep learning-based explanations for biomarker discovery.
(a) The DeepDream method presents random noise to a trained ANN model and iteratively adjusts this input to maximize the probability of the SZ class, resulting in a prototypical example of the SZ class according to the model. (b) After the model predicts the SZ class for a particular example, the output score of the model can be backpropagated through the layers by following one of several saliency or relevance propagation methods, resulting in a heatmap at the input that quantifies the contributions of each region/variable to the observed output. (c) By quantifying the increase in the prediction error caused by a small occlusion centered at different regions/variables of the input, an occlusion map can be generated that indicates the most relevant regions/variables. Although appealing for its simplicity, generating an occlusion map is computationally expensive, is sensitive to the occlusion size, and ignores interactions between different regions/variables. (d) The architecture of an ANN model can include attention layers with multiple attention heads focusing on different parts or elements of the input, making the ANN model more interpretable by design. For a given prediction, attention coefficients can be visualized to observe the attended (relevant) regions/variables of the input.

2013).

Two diagnosis studies employed approaches that depart significantly from the ones discussed thus far. These studies used approaches that take into account both the nonlinear aspects of DNNs and the interactions between input variables. The first study used a deep belief network (DBN), a special type of ANN that, after training in a classification task, can generate prototypical examples of the classes (SZ or healthy). The comparison between prototypes belonging to different classes revealed variables that were interpreted as relevant for differentiating the classes (Pinaya et al., 2016). An important aspect of this approach is that it takes advantage of the combinations of features or implicit statistical models that, through training, ANNs build and use for classifying cases. The models consist of combinations of features that were extracted and selected for their value for differentiating the classes, accounting for any potential interactions between the features. Thus, the prototypical examples generated from the internal models could be especially informative for discovering novel diagnostic-relevant differences between classes. While generative networks such as DBNs and generative adversarial networks (GANs) are capable of generating class prototypes directly due to special architectural designs, techniques have been developed that allow users to extract class prototypes out of trained regular, non-generative DNN classifiers. In general terms, the latter techniques, which include DeepDream (Fig. 6a) (Mordvintsev et al., 2015; Simonyan et al., 2013), use backpropagation to generate a configuration of input values (a prototype) that maximizes the probability of being assigned to a target class (e.g., SZ).

The second study employed a class saliency visualization approach to a trained convolutional DNN classifier (Oh et al., 2019). This type of approach uses backpropagation-like methods to assign weights to the input variables (e.g., voxels) representing the contribution of each input to a network's classification decision (diagnosis) in the specific context (set of interrelations) provided by the other inputs (Fig. 6b). The study used a simple version of this technique that estimates an input's relevance using the gradient of the DNNs output (diagnosis) with respect to the input (Simonyan et al., 2013). While appealing for its simplicity, this particular technique is known to be noisy and, sometimes, unreliable (Ancona et al., 2017). For this reason, more sophisticated and robust backpropagation-based methods for relevance maps have been proposed that we encourage the reader to explore (Ancona et al., 2017).

The studies discussed above are a valuable first step into using deep learning to obtain novel biomarkers and testable hypotheses to be assessed using experimental designs. Furthermore, future studies should look for differences between psychiatric conditions to identify biomarkers that are specific to SZ, discarding the possibility that they reflect mental illness in general.

To the best of our knowledge, the enormous potential offered by post hoc analysis methods in subtyping SZ remains unexplored. In classification tasks, the members of a class may have a large variability in their defining traits (e.g., the class of timekeeping devices will include images of wristwatches, pocket watches, pendulum clocks, sundials, hourglasses, water clocks, etc.). Thus, well-performing trained DNNs can assign the same label to elements that may have little or nothing in common (e.g., images of an hourglass and a pocket watch). While there may be no features common to all the members of a class, subgroups of members of the class may be more homogeneous and share many features (e.g., 8-shaped contour shared by hourglasses and circle-shaped contour shared by pocket watches). In this case, post hoc methods applied to trained DNN classifiers could reveal that the set of features that were most relevant to classify some members of a class (e.g., 8-shaped contour for hourglasses), are different from the set of features used to classify other members of the same class (e.g., circle-shaped contour with hands for pocket watches). By learning to assign umbrella term-labels to inhomogeneous elements, DNN classifiers could learn to extract the sets of input variables that characterize the different subgroups of members belonging to a class. In the case of SZ research, the exploration of high performance DNN classifiers trained on large, diverse datasets can provide curated, theory-agnostic sets of input variables (e.g., fMRI connectivity patterns) that could be used for subtyping and better understanding the rich variability hiding under the blanket term of SZ.

The field of explainable AI techniques is an active area of research, and each year more techniques are introduced, challenging the notion that DNNs are simple “black boxes”. Several challenges remain, as current explanations are still limited to highlighting the inputs associated with a specific output. Thus, they are insufficient to understand the complex input/output relationships modeled by DNNs (Darwiche, 2018).

4.3. Safety and knowledge generation: the increasing role of uncertainty and causality

In healthcare and other areas involving the well-being of people, it is crucial to have a clear understanding of the *uncertainty* associated with a model's predictions. Uncertainty refers to the inability to anticipate if a prediction is right or wrong; thus, it is a measure of how trustworthy a model's predictions are. In general, ANN classifiers work by estimating the input's probability of belonging to each one of the possible classes and then assigning the input to the class with the highest probability (e.g., if the probabilities assigned to an MRI scan are 0.2 for healthy, 0.3 for bipolar, and 0.5 for SZ, the diagnosis would be SZ). While class probabilities reflect how *confident* a classifier is in its predictions, this confidence does not necessarily match the *uncertainty* of the classifier's predictions. In fact, it is common for standard ANN classifiers to be *miscalibrated*, that is, to be under- or overconfident compared to their real accuracy (Guo et al., 2017; Nixon et al., 2019). For example, a DNN plant classifier could be very confident (predicted probability greater than 0.9) in labeling images of poison hemlocks as wild carrots, despite being wrong and almost incapable of differentiating both plants (low accuracy).

Bayesian ANNs, i.e., the combination of ANNs with the principles of Bayesian probability theory, offer a rigorous mathematical framework for quantifying uncertainty and, therefore, addressing the miscalibration problem affecting the implementation of ANNs. In this context, uncertainty can be divided in two types: aleatoric and epistemic uncertainty (Jospin et al., 2020). *Aleatoric uncertainty* results from uncontrollable random fluctuations in the data that arise from causes such as aleatoric changes in sensor sensitivity. This type of uncertainty is an intrinsic property of the data and cannot be reduced. Instead, *epistemic uncertainty* arises from a lack of knowledge about the system or problem that is being modeled due to insufficient data. Incomplete information, i.e., “holes” in the knowledge contained in the data, can increase prediction uncertainty when predictions are made near or inside the “holes” (e.g., how will a DNN classifier trained on adult patients perform on pediatric patients?). This type of uncertainty can be reduced by increasing the amount and information content of the training data and can be estimated using Bayesian ANN methods.

The process of training a non-Bayesian ANN consists in searching for a set of parameters (weights) that maximizes task performance. Training concludes with a point estimate of the parameters, usually by selecting the best set of parameters among those found during the search. Differently from this, in the Bayesian framework the objective is no longer to find a single parameter set for a task, but the full probability distribution of possible parameter sets given the training data. This probability distribution of models (one model per set of parameters) can be used to estimate the Bayesian model average, which is a weighted average of all the models' predictions (Wilson, 2020). This average reflects the disagreement between the predictions of different models and, compared to class probabilities, provides a better estimate of how (epistemically) uncertain or trustworthy a prediction really is.

Although most Bayesian methods are still not easily applicable to large scale DNNs, Bayesian ANNs are an active area of research in which novel techniques (Gal and Ghahramani, 2016) are slowly reducing the high computational burden (e.g., training several ANNs with different parameters) associated with estimating parameter distributions (Jospin et al., 2020). This type of research is likely to have a tremendous impact on the adoption of DNNs in the clinic, where miscalibrated (over- and underconfident) AI tools can become an unacceptable safety risk. This will likely benefit patients receiving AI-informed diagnosis and tailored treatment interventions.

As mentioned, epistemic uncertainty can be reduced by increasing the amount of training data. This approach has worked well in areas such as natural language processing and computer vision, which have easy online access to massive datasets with billions of training examples (Brown et al., 2020a). However, in areas such as SZ research, this brute-force approach to reduce uncertainty is unfeasible and, therefore, it is necessary to look for alternative methods capable of harnessing the available data more efficiently.

Judea Pearl, a pioneer in ML causality and the creator of Bayesian networks, has pointed out that the enormous power of the DNNs that are driving the ongoing AI revolution comes, almost entirely, from their capacity to learn complex input/output *statistical associations* (Pearl, 2018, 2020). As discussed before (Section 2.2), DNNs learn these associations by minimizing the error in predicting the output that is most likely to *co-occur* with each input in an enormous collection of input/output pairs (training dataset). While this data-driven, co-occurrence-learning strategy has succeeded in solving a wide range of prediction and classification tasks, it is a slow and data-inefficient strategy that, in many aspects, is analogous to the natural selection process driving Darwinian evolution (Pearl, 2018, 2020). Differently from this, human learning and problem solving is an active process that relies on the development and manipulation of explanatory *causal models* (hypotheses) of reality (problems) that are tested against data, specially selected/collected for the purpose of falsifying or supporting the model (high data-efficiency). A causal model is a set of hypotheses that can explain the statistical associations in the data in terms of *causal relationships*; that is, relationships between couples of sequential events in which the first event of each couple, the cause, is a part of the mechanism that explains or generates the second event of the couple, the effect (e.g., a question elicits [causes] an answer or adrenaline induces [causes] changes in heart rate).

Compared to association-learning systems, causal systems are fast and efficient (e.g., humans do not need to review tens of thousands of checkers matches to learn how to play), and have allowed humans, assisted by technology, to “evolve” at a super-evolutionary pace (Pearl, 2018). For example, while it took hundreds of millions of years of natural selection for birds to evolve the capacity to fly after the first feathers appeared, it took just a few decades for humans to build the first supersonic jet after the Wright brothers built the first heavier-than-air, motor-operated, aircraft. Furthermore, separate from purely statistical models, causal models allow to accurately predict the behavior of a system in situations that have never been encountered before in the training data (e.g., black holes were derived from the equations of general relativity decades before they were experimentally confirmed). Finally, causal models are the only ones capable of pursuing what is arguably the main goal of science, to find the reason *why*, i.e. the *causal mechanisms* underlying statistical associations (Castro et al., 2020; Pearl and Mackenzie, 2018).

According to Pearl, in order to move the field of AI forward and reduce the gap between humans and machines in learning speed, data use efficiency, and problem-solving capacity, it is necessary to equip learning machines with *causal reasoning* tools (Pearl, 2018, 2020). That is, with mechanisms that allow machines to generate *causal models* (hypotheses) of problems (e.g., treatment selection), use the models to anticipate (predict) the effects of interventions (tests), compare the models' predictions with data specially selected or collected to test the models, and then actively decide the right course of action (e.g., refine the model, further interventions, collection of more data, etc.) (Pearl, 2018; Pearl and Mackenzie, 2018). This type of causal model-oriented AI systems will underlie what has been called the third wave of AI, with the first wave being traditional machine learning (handcrafted knowledge), and the second, the ongoing AI revolution (statistical learning) (Launchbury, 2017). While causal AI is still in its early infancy, it is attracting the attention of on an increasing number of researchers that have started to work in the area (Bengio et al., 2019; Ke et al., 2019; Nauta et al., 2019). Until causality is fully embedded in deep learning models, insights related to the mechanisms of SZ extracted from DNNs must be assessed by researchers and followed up by empirical tests to discard spurious or irrelevant correlations (e.g., people take aspirin when they have a cold, thus, to prevent colds, aspirins should be avoided).

5. Discussion

In this article, we presented an overview of the existing literature on deep learning methods applied to SZ research. In general, existing studies have yielded impressive results in terms of accuracy in classification and outcome prediction tasks, justifying the increasing interest in deep learning approaches. However, methodological issues affecting the generalizability of the results in several studies suggest that some reports may be overoptimistic and should be taken with caution. The most relevant issues are the small size of the samples used for developing DNN models, and the lack of independence between the testing dataset and the training and validation datasets.

When the training, validation, and testing datasets come from a small sample of subjects studied under specific experimental conditions (e.g., EEG data collected by a research group using the same EEG device and experimental setting), it is unlikely that these datasets will be representative of the rich variability of data collected under different experimental conditions (e.g., EEG data collected by a different research group using a different EEG system and experimental setting). Thus, it is unlikely that the results obtained on small datasets will generalize appropriately to the set of independent datasets. For example, diagnostic models trained on samples of 100–200 examples or less (a very small sample size for DNN training) usually reached accuracies >90%, while models trained on large multi-site datasets reached lower accuracies of ~80% (Yan et al., 2019; Yang et al., 2019; Zeng et al., 2018; Zhao et al., 2020).

Current state-of-the-art deep learning models are trained on samples ranging between hundreds of thousands and hundreds of millions of examples (Devlin et al., 2018). It is unlikely that datasets with similar sample sizes will be available any time soon for training deep learning models in the context of SZ research. However, while this amount of *real* data is currently out of reach, it is possible to increase the size of datasets by complementing *real* data with *synthetic* data generated with data augmentation techniques. Data augmentation includes techniques such as simulating data with physical or empirical models, generating data with generative adversarial networks (GANs), or modifying real data with transformations that preserve class assignment or are irrelevant for the task (e.g., rotating a headshot does not change the identity of the person depicted in it) (Lashgari et al., 2020; Shorten and Khoshgoftaar, 2019). Moreover, novel algorithms have allowed researchers to use data augmentation to pre-train models on unlabeled data (e.g., fMRI from undiagnosed people) (Chen et al., 2020) that are usually easier to find than data from specific clinical populations. To our knowledge, only one study has used advanced data augmentation techniques to increase sample size while enriching the variability (diversity) of small datasets in SZ research (Niu et al., 2019).

Another option to increase sample sizes would be to train models on multimodal data. Besides the improvements in performance resulting from training models on cases with data from multiple modalities, models designed for multimodal data have the advantage that they can be trained also on partial data, i.e., data that are missing one of the modalities (Guo et al., 2019). Thus, single-modality datasets can be combined and used to train the same model. Each modality will contribute to shape the high-level features of these models. Very few studies have used multimodal data to train deep learning models in SZ research (Li et al., 2020; Salvador et al., 2019).

The independence of the testing dataset is one of the most basic requirements for developing any ML model and it is crucial for obtaining unbiased estimations of the model's generalizability. Several studies compromised the independence of the testing data by informing decisions about the model's hyperparameters and/or data preprocessing steps using results obtained on the testing dataset. A common mistake was to use cross-validation (single loop) instead of nested cross-validation (with an outer loop for testing and an inner loop for validation) for both validation (e.g., hyperparameter selection and model design) and testing. Furthermore, few studies estimated their model's generalizability to data coming from a different population or acquired using different protocols (external dataset), where performance might drop significantly. For example, in a multi-site study, the accuracy of a classification model fell from 88.6% ($n = 866$, IDT—I, 10-fold CV) to 70% ($n = 60$, ODT—I) when testing on a dataset collected from a held-out site (Oh et al., 2020). Other multi-site studies also showed accuracy drops ranging from 1.4% to 18.5% (see Fig. 4f) when switching from pooled-sites cross-validation (i.e., data from all sites are combined in a single dataset) to leave-one-site-out cross-validation (i.e., on each iteration, all the data collected at one of the study sites were used exclusively for testing), falling from 81–85% ($n = 149$ to 1100, IDT-NI, 5-fold or 10-fold CV) to 62.5–81% ($n = 734$ to 1100, ODT-NI, 5 or 7 sites) (Lei et al., 2019; Yan et al., 2019; Zeng et al., 2018; Zhao et al., 2020).

5.1. Moving forward

In most of the studies discussed here, deep learning models were trained to use brain imaging data (e.g., fMRI) to differentiate between SZ and healthy controls. While this is a valuable first step, these models have limited utility for clinicians who usually can easily distinguish between SZ patients and healthy people. Instead, it would be more relevant to focus on developing deep learning systems aimed at distinguishing SZ from other psychiatric disorders with the purpose of aiding clinicians in the selection of treatment interventions, predicting clinical outcomes (e.g., the probability of a first episode of SZ) to implement prophylactic interventions, or predicting a patient's response to medication to select the most effective pharmacological treatment that has the least side effects (personalized medicine). While a few studies have addressed some of these issues (Ahmedt Aristizabal et al., 2020; Fernando et al., 2020; Rezaii et al., 2019), urgent clinical needs remain unmet. For instance, we found no studies focused on predicting episodes of clinical decompensation in established SZ patients, despite decompensation remaining a major clinical problem (Robinson et al., 1999; Zipursky et al., 2014). Redirecting research efforts to solving real problems encountered in clinical practice has the potential of improving treatment efficacy and relieving patients and their caregivers from some of the burden associated with SZ.

5.2. Future directions and ethical AI development, deployment, and use

One could argue that the onset of the ongoing deep learning revolution can be traced back to the creation of ImageNet in the late 2000s (Deng et al., 2009), a large, publicly available dataset of labeled images designed for ML projects. The open access to a common benchmark and the annual competition that followed, boosted the field of computer vision by providing an objective way for assessing and, most importantly, comparing the performance of different ML approaches designed to identify the objects in ImageNet's images. In the field of SZ research, there is an urgent need for analogous large, easily accessible, multi-diagnosis datasets and agreed evaluation metrics that allow researchers to assess and compare ML approaches. We have no doubt that these resources will accelerate the development of more powerful and accurate ML models for SZ research. However, differently from what happened with ImageNet and other datasets, there are serious ethical issues associated with using data collected from humans for developing ML systems that need to be addressed.

The first issue refers to the ***ethical collection and release of data***: safeguards should be implemented to guarantee that the collection of data from SZ patients and other vulnerable populations respects the basic rights of the people from whom the data were collected. One such safeguard could be to ensure, for example, that the protocols used for data collection and sharing are reviewed and approved by a qualified institutional committee entrusted with the protection of human subjects. Furthermore, psychiatric diagnoses could be used for discriminatory practices by employers, insurance companies, and others, thus, it is crucial to take measures (e.g., stripping the data from personal identifiable information) to ensure the privacy of the people contributing to the dataset.

The second, refers to ***ethical management of biased and unbalanced data***: A bias can be understood as an unfair (unjustified) tendency to have a consistently favorable or unfavorable perception, attitude, or behavior towards individuals sharing a sensitive attribute such as race, gender, place of birth, religion, or social class. Biases are ubiquitous among individuals, institutions, and societies, therefore, along with useful associations (e.g., relationship between brain connectivity and SZ), datasets collected from humans will carry a raft of unfair associations (biases) as well (e.g., patients with African ancestry have a higher risk of receiving a misdiagnosis of SZ (Akinhanmi et al., 2018)). Furthermore, sensitive attributes such as race (e.g., African and Hispanic vs European ancestry) and gender (e.g., LGBT vs heterosexual) are not uniformly distributed among the population, thus, randomly collected datasets will replicate the attribute unbalance and, therefore, are very likely to under-represent the characteristics of minorities and marginalized groups in comparison to the hegemonic group/culture.

ML systems learn the statistical patterns and associations in the data that are useful for solving the task at hand, and apply them to generate their outputs. Thus, unless special remediation mechanisms are put in place (as discussed next), biased and unbalanced datasets will lead to biased and selectively inaccurate ML systems. For instance, a DNN diagnosis system trained on brain connectivity measures collected on a large sample of mostly male SZ patients and controls, will likely disregard gender differences (Ingalhalikar et al., 2014; Li et al., 2016) and use male-derived patterns to classify both male and female patients. This situation may not only increase the risk of misdiagnosing females but also mischaracterize the neural underpinnings of female SZ patients. In fact, post hoc techniques applied to this classifier (see Section 4.2) would reveal biomarkers optimized for classifying male SZ patients that, by disregarding biases and unbalanced samples, would be attributed to female patients as well, hiding any potential gender differences. A dramatic example of the real-life consequences that could follow from algorithmic biases is what happened with a widely used commercial algorithm used for identifying patients with complex conditions that will benefit from receiving extra care. It was found that at a given healthcare need score, black patients were considerably sicker than white patients. Furthermore, it was revealed that, without this disparity, the percentage of black patients that received extra care would have increased from 17.7 to 46.5% (Obermeyer et al., 2019).

During the last few years, ML fairness, i.e., the guarantee that sensitive attributes in the data do not affect the output of an ML algorithm in an unfair way, has become a highly active and increasingly growing area of research. As discussed above, disregarding ML fairness may seriously affect the well-being of patients and may hinder our understanding of the mechanisms underlying SZ. Thus, while reviewing the rich literature on ML fairness goes beyond the scope of this article, we provide a broad description of the classes of methods that can be used for ensuring ML fairness in SZ research and refer interested readers to the following revisions of the literature (Chiappa and Isaac, 2018; Chouldechova and Roth, 2018; Oneto and Chiappa, 2020). In general terms, methods for ensuring ML fairness can be grouped in three classes, depending on what stage of the developing process of an ML algorithm they are applied (Oneto and Chiappa, 2020). The first class of methods are applied to the data to remove the biases before training the ML algorithm. The second class of methods are applied on the training process to force the ML algorithm to produce fair outputs despite being trained on biased or unbalanced data. Finally, the third class of methods are applied to the biased outputs of trained ML algorithms to correct them towards a fair (unbiased) output.

The third and last issue refers to the ***ethical use of algorithms***: ML algorithms trained for diagnosis, prediction of clinical outcomes, biomarker discovery, and related uses can improve patient diagnosis, boost personalized medicine, and lead to novel discoveries. However, the same algorithms could be easily adapted for different purposes such as screening candidates for a job, calculating insurance fees, targeted advertising, and student selection and admission, potentially leading to discriminatory practices that have a negative impact on the very people that these technologies are supposed to help. Thus, as stated by Joseph Redmon, the creator of the highly influential (~16,000 citations) *You look only once* (YOLO) algorithm for computer vision (Redmon et al., 2016), "*as researchers we have a responsibility to at least consider the harm our work might be doing and think of ways to mitigate it. We owe the world that much*" (Redmon and Farhadi, 2018).

The AI revolution that has been occurring during the last few years, is changing the way in which we do SZ research. Our ability to extract useful information from high-dimensional and multimodal data has never been better. Likewise, the accuracy of automatic data-based diagnosis and outcome prediction systems is higher than ever. These tools promise to increase our capacity for implementing prophylactic or more effective personalized treatment interventions. While these methods can help us to ease some of the burden of SZ from patients and caregivers, they may bring a series of ethical problems that need to be carefully assessed and addressed. In addition, considering the negative impact that a misdiagnosis or erroneous prediction may have on someone's life, deep learning-powered systems will require continuous error assessment and suitable protocols to safeguard the well-being of people.

CRediT authorship contribution statement

JAC, NIT, and PE were involved in writing the manuscript and conducting the literature review. DCD contributed with writing and editing the manuscript. JAC and NIT contributed equally to the manuscript.

Declaration of competing interest

None.

Acknowledgment

The authors would like to acknowledge the invaluable help of Leigh Taylor Flynn during the preparation of this manuscript. JAC acknowledges the support of the Brain and Behavior Research Foundation through its Young Investigator Award (23848) and PE acknowledges support from ANID-Chile through FONDECYT grant (1171678).

Role of the funding source

JAC reports receiving funding administered through Yale University from the National Institutes of Health, and the Brain and Behavior Research Foundation. DCD reports receiving funding administered through Yale University from the National Institutes of Health, the Brain and Behavior Research Foundation, the Wallace Foundation, CH TAC, LLC, Takeda, Roche, and the Heffter Research Institute. PE reports receiving funding from ANID-Chile.

References

Ahmedt Aristizabal, D., Fernando, T., Denman, S., Robinson, J.E., Sridharan, S., Johnston, P.J., Laurens, K.R., Fookes, C., 2020. Identification of children at risk of schizophrenia via deep learning and EEG responses. IEEE J. Biomed. Health Inform. 1–7.

Akinhanmi, M.O., Biernacka, J.M., Strakowski, S.M., McElroy, S.L., Balls Berry, J.E., Merikangas, K.R., Assari, S., McInnis, M.G., Schulze, T.G., LeBoyer, M., 2018. Racial disparities in bipolar disorder treatment and research: a call to action. Bipolar Disord. 20 (6), 506–514.

Alnæs, D., Kaufmann, T., van der Meer, D., Córdova-Palomera, A., Rokicki, J., Moberget, T., Bettella, F., Agartz, I., Barch, D.M., Bertolino, A., Brandt, C.L., Cervenka, S., Djurovic, S., Doan, N.T., Eisenacher, S., Fatouros-Bergman, H., Flyckt, L., Di Giorgio, A., Haatveit, B., Jönsson, E.G., Kirsch, P., Lund, M.J., Meyer-Lindenberg, A., Pergola, G., Schwarz, E., Smeland, O.B., Quarto, T., Zink, M., Andreassen, O.A., Westlye, L.T., Consortium, f.t.K.S.P., 2019. Brain heterogeneity in schizophrenia and its association with polygenic risk. JAMA Psychiatry 76 (7), 739–748.

Ancona, M., Ceolini, E., Öztireli, C., Gross, M., 2017. Towards better understanding of gradient-based attribution methods for deep neural networks. arXiv preprint arXiv: 1711.06104.

Anumanchipalli, G.K., Chartier, J., Chang, E.F., 2019. Speech synthesis from neural decoding of spoken sentences. Nature 568 (7753), 493–498.

Apicella, A., Donnarumma, F., Isgrò, F., Prevete, R., 2021. A survey on modern trainable activation functions. Neural Netw. 138, 14–32.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M.X., Cao, Y., Foster, G., Cherry, C., 2019. Massively multilingual neural machine translation in the wild: findings and challenges. arXiv preprint arXiv:1907.05019.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., 2020. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.

Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. Advances in neural information processing systems 33.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Barnes, T.R., Leeson, V.C., Mutsatsa, S.H., Watt, H.C., Hutton, S.B., Joyce, E.M., 2008. Duration of untreated psychosis and social function: 1-year follow-up study of first-episode schizophrenia. Br. J. Psychiatry 193 (3), 203–209.

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., Vardoulakis, L.M., 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy, proceedings of the 2020 CHI conference on human factors in computing systems, pp. 1–12.

Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., Pal, C., 2019. A meta-transfer objective for learning to disentangle causal mechanisms. arXiv preprint arXiv:1901.10912.

Bleuler, E., 1950. Dementia praecox. International Universities Press, New York.

Bowen, E.F.W., Burgess, J.L., Granger, R., Kleinman, J.E., Rhodes, C.H., 2019. DLPFC transcriptome defines two molecular subtypes of schizophrenia. Transl. Psychiatry 9 (1), 147.

Brinker, T.J., Hekler, A., Enk, A.H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Holland-Letz, T., Utikal, J.S., von Kalle, C., Ludwig-Peitsch, W., Sirokay, J., Heinzerling, L., Albrecht, M., Baratella, K., Bischof, L., Chorti, E., Dith, A., Drusio, C., Giese, N., Gratsias, E., Griewank, K., Hallasch, S., Hanhart, Z., Herz, S., Hohaus, K., Jansen, P., Jockenhöfer, F., Kanaki, T., Knispel, S., Leonhard, K., Martaki, A., Matei, L., Matull, J., Olischewski, A., Petri, M., Placke, J.-

M., Raub, S., Salva, K., Schlott, S., Sody, E., Steingrube, N., Stoffels, I., Ugurel, S., Zaremba, A., Gebhardt, C., Booken, N., Christolouka, M., Buder-Bakhaya, K., Bokor-Billmann, T., Enk, A., Gholam, P., Hänßle, H., Salzmann, M., Schäfer, S., Schäkel, K., Schank, T., Bohne, A.-S., Deffaa, S., Drerup, K., Egberts, F., Erkens, A.-S., Ewald, B., Falkvoll, S., Gerdes, S., Harde, V., Hauschild, A., Jost, M., Kosova, K., Messinger, L., Metzner, M., Morrison, K., Motamedi, R., Pinczker, A., Rosenthal, A., Scheller, N., Schwarz, T., Stölzl, D., Thielking, F., Tomaschewski, E., Wehkamp, U., Weichenthal, M., Wiedow, O., Bär, C.M., Bender-Säbelkampf, S., Horbrügger, M., Karoglan, A., Kraas, L., Faulhaber, J., Geraud, C., Guo, Z., Koch, P., Linke, M., Maurier, N., Müller, V., Thomas, B., Utikal, J.S., Alamri, A.S.M., Baczako, A., Berking, C., Betke, M., Haas, C., Hartmann, D., Heppt, M.V., Kilian, K., Krammer, S., Lapczynski, N.L., Mastnik, S., Nasifoglu, S., Ruini, C., Sattler, E., Schlaak, M., Wolff, H., Achatz, B., Bergbreiter, A., Drexler, K., Ettinger, M., Haferkamp, S., Halupczok, A., Hegemann, M., Dinauer, V., Maagk, M., Mickler, M., Philipp, B., Wilm, A., Wittmann, C., Gesierich, A., Glutsch, V., Kahlert, K., Kerstan, A., Schilling, B., Schrüfer, P., 2019. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur. J. Cancer 113, 47–54.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., 2020a. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020b. Language Models Are Few-Shot Learners.

Brugger, S.P., Howes, O.D., 2017. Heterogeneity and homogeneity of regional brain structure in schizophrenia: a meta-analysis. JAMA Psychiatry 74 (11), 1104–1111.

Calhas, D., Romero, E., Henriques, R., 2020. On the use of pairwise distance learning for brain signal classification with limited observations. Artif. Intell. Med. 105, 101852.

Castro, D.C., Walker, I., Glocker, B., 2020. Causality matters in medical imaging. Nat. Commun. 11 (1), 1–10.

Chen, J., Wu, J.s., Mize, T., Shui, D., Chen, X., 2018. Prediction of schizophrenia diagnosis by integration of genetically correlated conditions and traits. J. NeuroImmune Pharmacol. 13, 532–540.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A Simple Framework for Contrastive Learning of Visual Representations.

Chiappa, S., Isaac, W.S., 2018. A causal bayesian networks viewpoint on fairness, IFIP international summer school on privacy and identity management. Springer, pp. 3–20.

Chouldechova, A., Roth, A., 2018. The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810.

Dakka, J., Bashivan, P., Gheiratmand, M., Rish, I., Jha, S., Greiner, R., 2017. Learning Neural Markers of Schizophrenia Disorder Using Recurrent Neural Networks (NeurIPS).

Darwiche, A., 2018. Human-level intelligence or animal-like abilities? Commun. ACM 61 (10), 56–67.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1, 4171–4186.

Elsken, T., Metzen, J.H., Hutter, F., 2018. Neural architecture search: a survey. arXiv preprint arXiv:1808.05377.

Farmer, A.E., McGuffin, P., Spitznagel, E.L., 1983. Heterogeneity in schizophrenia: a cluster-analytic approach. Psychiatry Res. 8 (1), 1–12.

Fernando, T., Denman, S., Ahmedt-Aristizabal, D., Sridharan, S., Laurens, K.R., Johnston, P., Fookes, C., 2020. Neural memory plasticity for medical anomaly detection. Neural Netw. 127, 67–81.

Fusar-Poli, P., Bonoldi, I., Yung, A.R., Borgwardt, S., Kempton, M.J., Valmaggia, L., Barale, F., Caverzasi, E., McGuire, P., 2012. Predicting psychosis: meta-analysis of transition outcomes in individuals at high clinical risk. Arch. Gen. Psychiatry 69 (3), 220–229.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: representing model uncertainty in deep learning, international conference on machine learning. PMLR, pp. 1050–1059.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks, proceedings of the fourteenth international conference on artificial intelligence and statistics, pp. 315–323.

Guest, P.C., Martins-de-Souza, D., Schwarz, E., Rahmoune, H., Alsaif, M., Tomasik, J., Turck, C.W., Bahn, S., 2013. Proteomic profiling in schizophrenia: enabling stratification for more effective treatment. Genome Med. 5 (3), 25.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks, international conference on machine learning. PMLR, pp. 1321–1330.

Guo, W., Wang, J., Wang, S., 2019. Deep multimodal representation learning: a survey. IEEE Access 7, 63373–63394.

Holderness, E., Miller, N., Cawkwell, P., Bolton, K., Meteer, M., Pustejovsky, J., Hall, M. H., 2019. Analysis of risk factor domains in psychosis patient health records. J. Biomed. Semant. 10, 19.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Netw. 2 (5), 359–366.

Hubel, D.H., Wiesel, T.N., 1968. Receptive fields and functional architecture of monkey striate cortex. J. Physiol. 195 (1), 215–243.

Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T.D., Elliott, M.A., Ruparel, K., Hakonarson, H., Gur, R.E., Gur, R.C., Verma, R., 2014. Sex differences in the structural connectome of the human brain. Proc. Natl. Acad. Sci. 111 (2), 823–828.

Jospin, L.V., Buntine, W., Boussaid, F., Laga, H., Bennamoun, M., 2020. Hands-on Bayesian neural networks—a tutorial for deep learning users. arXiv preprint arXiv: 2007.06823.

Ke, N.R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Schölkopf, B., Mozer, M.C., Pal, C., Bengio, Y., 2019. Learning neural causal models from unknown interventions. arXiv preprint arXiv:1910.01075.

Kim, J., Calhoun, V.D., Shim, E., Lee, J.H., 2016. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. NeuroImage 124, 127–146.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kormilitzin, A., Vaci, N., Liu, Q., Nevado-Holgado, A., 2020. Med7: A Transferable Clinical Natural Language Processing Model for Electronic Health Records.

Lashgari, E., Liang, D., Maoz, U., 2020. Data augmentation for deep-learning-based electroencephalography. J. Neurosci. Methods 108885.

Launchbury, J., 2017. A DARPA Perspective on Artificial Intelligence.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Lee, M., Sehatpour, P., Hoptman, M.J., Lakatos, P., Dias, E.C., Kantrowitz, J.T., Martinez, A.M., Javitt, D.C., 2017. Neural mechanisms of mismatch negativity dysfunction in schizophrenia. Mol. Psychiatry 22 (11), 1585–1593.

Lei, D., Pinaya, W.H.L., Van Amelsvoort, T., Marcelis, M., Donohoe, G., Mothersill, D.O., Corvin, A., Gill, M., Vieira, S., Huang, X., Lui, S., Scarpazza, C., Young, J., Arango, C., Bullmore, E., Qiyong, G., McGuire, P., Mechelli, A., 2019. Detecting schizophrenia at the level of the individual: relative diagnostic value of whole-brain images, connectome-wide functional connectivity and graph-based metrics. Psychological medicine, 1–10.

Li, R., Ma, X., Wang, G., Yang, J., Wang, C., 2016. Why sex differences in schizophrenia? J. Transl. Neurosci. 1 (1), 37.

Li, G., Han, D., Wang, C., Hu, W., Calhoun, V.D., Wang, Y.-P., 2020. Application of deep canonically correlated sparse autoencoder for the classification of schizophrenia. Comput. Methods Prog. Biomed. 183, 105073.

Liang, S.G., Greenwood, T.A., 2015. The impact of clinical heterogeneity in schizophrenia on genomic analyses. Schizophr. Res. 161 (2–3), 490–495.

Lipton, Z.C., 2018. The mythos of model interpretability. Queue 16 (3), 31–57.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M., 2020. Deep learning for generic object detection: a survey. Int. J. Comput. Vis. 128 (2), 261–318.

Ma, K.W.-D., Lewis, J., Kleijn, W.B., 2020. The HSIC Bottleneck: Deep Learning Without Back-Propagation. AAAI, pp. 5085–5092.

Marshall, M., Lewis, S., Lockwood, A., Drake, R., Jones, P., Croudace, T., 2005. Association between duration of untreated psychosis and outcome in cohorts of first-episode patients: a systematic review. Arch. Gen. Psychiatry 62 (9), 975–983.

Matsubara, T., Tashiro, T., Uehara, K., 2019. Deep neural generative model of functional MRI images for psychiatric disorder diagnosis. IEEE Trans. Biomed. Eng. 66 (10), 2768–2779.

McCarthy, J., 1983. The little thoughts of thinking machines. Psychol. Today 17 (12), 46–49.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013b. Distributed representations of words and phrases and their compositionality, advances in neural information processing systems, pp. 3111–3119.

Miotto, R., Li, L., Kidd, B.A., Dudley, J.T., 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci. Rep. 6, 26094.

Mordvintsev, A., Olah, C., Tyka, M., 2015. Deepdream-a code example for visualizing neural networks. Google Res. 2 (5).

Naderi, H., Soleimani, B.H., Matwin, S., 2019. Multimodal Deep Learning for Mental Disorders Prediction From Audio Speech Samples (NeurIPS).

Nauta, M., Bucur, D., Seifert, C., 2019. Causal discovery with attention-based convolutional neural networks. Mach. Learn. Knowl. Extract. 1 (1), 312–340.

Niu, Y.W., Lin, Q.H., Qiu, Y., Kuang, L.D., Calhoun, V.D., 2019. Sample augmentation for classification of schizophrenia patients and healthy controls using ICA of fMRI data and convolutional neural networks, 10th international conference on intelligent control and information processing, ICICIP 2019. Institute of Electrical and Electronics Engineers Inc., pp. 297–302.

Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D., 2019. Measuring Calibration in Deep Learning (CVPR Workshops).

Nwankpa, C., Ijomah, W., Gachagan, A., Marshall, S., 2018. Activation functions: comparison of trends in practice and research for deep learning. arXiv preprint arXiv:1811.03378.

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366 (6464), 447–453.

Oh, K., Kim, W., Shen, G., Piao, Y., Kang, N.I., Oh, I.S., Chung, Y.C., 2019. Classification of schizophrenia and normal controls using 3D convolutional neural network and outcome visualization. Schizophr. Res. 212, 186–195.

Oh, J., Oh, B.-L., Lee, K.-U., Chae, J.-H., Yun, K., 2020. Identifying schizophrenia using structural MRI with a deep learning algorithm. Front. Psychiatry 11, 16.

Olah, C., Mordvintsev, A., Schubert, L., 2017. Feature visualization. Distill 2 (11), e7.

Oneto, L., Chiappa, S., 2020. Fairness in machine learning. In: Recent Trends in Learning From Data. Springer, pp. 155–196.

Patel, P., Aggarwal, P., Gupta, A., 2016. Classification of Schizophrenia Versus Normal Subjects Using Deep Learning. ACM International Conference Proceeding Series. Association for Computing Machinery, New York, pp. 1–6.

Pearl, J., 2018. Theoretical impediments to machine learning with seven sparks from the causal revolution. arXiv preprint arXiv:1801.04016.

Pearl, J., 2020. Radical empiricism and machine learning research. In: Causal Analysis in Theory and Practice (Blog). July 26.

Pearl, J., Mackenzie, D., 2018. The Book of Why: The New Science of Cause and Effect. Basic Books.

Phang, C.R., Noman, F., Hussain, H., Ting, C.M., Ombao, H., 2020. A multi-domain connectome convolutional neural network for identifying schizophrenia from EEG connectivity patterns. IEEE J. Biomed. Health Inform. 24 (5), 1333–1343.

Pinaya, W.H., Gadelha, A., Doyle, O.M., Noto, C., Zugman, A., Cordeiro, Q., Jackowski, A.P., Bressan, R.A., Sato, J.R., 2016. Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. Sci. Rep. 6, 38897.

Pinaya, W.H., Mechelli, A., Sato, J.R., 2019. Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: a large-scale multi-sample study. Hum. Brain Mapp. 40, 944–954.

Qureshi, M.N.I., Oh, J., Lee, B., 2019. 3D-CNN based discrimination of schizophrenia using resting-state fMRI. Artif. Intell. Med. 98, 10–17.

Redmon, J., Farhadi, A., 2018. Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: unified, real-time object detection, proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.

Rezaii, N., Walker, E., Wolff, P., 2019. A machine learning approach to predicting psychosis using semantic density and latent content analysis. NPJ Schizophr. 5, 9.

Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., de Berker, A., Ganguli, S., 2019. A deep learning framework for neuroscience. Nat. Neurosci. 22 (11), 1761–1770.

Robinson, D., Woerner, M.G., Alvir, J.M.J., Bilder, R., Goldman, R., Geisler, S., Koreen, A., Sheitman, B., Chakos, M., Mayerhoff, D., Lieberman, J.A., 1999. Predictors of relapse following response from a first episode of schizophrenia or schizoaffective disorder. Arch. Gen. Psychiatry 56 (3), 241–247.

Roy, A., 2012. A theory of the brain: localist representation is used widely in the brain. Front. Psychol. 3, 551.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536.

Salvador, R., Canales-Rodríguez, E., Guerrero-Pedraza, A., Sarró, S., Tordesillas-Gutiérrez, D., Maristany, T., Crespo-Facorro, B., McKenna, P., Pomarol-Clotet, E., 2019. Multimodal integration of brain images for MRI-based diagnosis in schizophrenia. Front. Neurosci. 13, 1203.

Schäfer, A.M., Zimmermann, H.-G., 2007. Recurrent neural networks are universal approximators. Int. J. Neural Syst. 17 (04), 253–263.

Senior, M., Burghart, M., Yu, R., Kormilitzin, A., Liu, Q., Vaci, N., Nevado-Holgado, A., Pandit, S., Zlodre, J., Fazel, S., 2020. Identifying predictors of suicide in severe mental illness: a feasibility study of a clinical prediction rule (Oxford Mental Illness and Suicide Tool or OxMIS). Front. Psychiatry 11, 268.

Shickel, B., Heesacker, M., Benton, S., Rashidi, P., 2017. Hashtag Healthcare: From Tweets to Mental Health Journals Using Deep Transfer Learning.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6 (1), 60.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., 2016. Mastering the game of Go with deep neural networks and tree search. Nature 529 (7587), 484–489.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., Schmidt, L., 2020. Measuring robustness to natural distribution shifts in image classification. Adv. Neural Inf. Proces. Syst. 33.

Teney, D., Kafle, K., Shrestha, R., Abbasnejad, E., Kanan, C., Hengel, A.v.d., 2020. On the Value of Out-of-Distribution Testing: An Example of Goodhart's Law. arXiv preprint arXiv:2005.09241.

Vieira, S., Gong, Q.-y., Pinaya, W.H.L., Scarpazza, C., Tognin, S., Crespo-Facorro, B., Tordesillas-Gutierrez, D., Ortiz-García, V., Setien-Suero, E., Scheepers, F.E., Van Haren, N.E.M., Marques, T.R., Murray, R.M., David, A., Dazzan, P., McGuire, P., Mechelli, A., 2020. Using machine learning and structural neuroimaging to detect first episode psychosis: reconsidering the evidence. Schizophr. Bull. 46, 17–26.

Voineskos, A.N., 2015. Genetic underpinnings of white matter 'connectivity': heritability, risk, and heterogeneity in schizophrenia. Schizophr. Res. 161 (1), 50–60.

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C.P., Clarke, D., Gu, M., Emani, P., Yang, Y.T., Min, X., Gandal, M.J., Lou, S., Zhang, J., Park, J.J., Yan, C., KyongRhie, S., Manakongtreecheep, K., Zhou, H., Aparna Natha, A., Peters, M., Mattei, E., Fitzgerald, D., Brunetti, T., Moore, J., Jiang, Y., Girdhar, K., Hoffman, G. E., Kalayci, S., Gümüş, Z.H., Crawford, G.E., Roussos, P., Akbarian, S., Jaffe, A.E., White, K.P., Weng, Z., Sestan, N., Geschwind, D.H., Knowles, J.A., Gerstein, M.B., 2018. Comprehensive functional genomic resource and integrative model for the human brain. Science 362, 1266.

Wang, T., Bezerianos, A., Cichocki, A., Li, J., 2019. Multi-Kernel Capsule Network for Schizophrenia Identification.

Wilson, A.G., 2020. The case for Bayesian deep learning. arXiv preprint arXiv:2001.10995.

Xu, J., Li, H., Zhou, S., 2015a. An overview of deep generative models. IETE Tech. Rev. 32 (2), 131–139.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015b. Show, attend and tell: neural image caption generation with visual attention, international conference on machine learning, pp. 2048–2057.

Yan, W., Calhoun, V., Song, M., Cui, Y., Yan, H., Liu, S., Fan, L., Zuo, N., Yang, Z., Xu, K., Yan, J., Lv, L., Chen, J., Chen, Y., Guo, H., Li, P., Lu, L., Wan, P., Wang, H., Wang, H., Yang, Y., Zhang, H., Zhang, D., Jiang, T., Sui, J., 2019. Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site FMRI data. EBioMedicine 47, 543–552.

Yang, B., Chen, Y., Shao, Q.-M., Yu, R., Li, W.-B., Guo, G.-Q., Jiang, J.-Q., Pan, L., 2019. Schizophrenia classification using fMRI data based on a multiple feature image capsule network ensemble. IEEE Access 7, 109956–109968.

Zela, A., Klein, A., Falkner, S., Hutter, F., 2018. Towards automated deep learning: efficient joint neural architecture and hyperparameter search. arXiv preprint arXiv:1807.06906.

Zeng, L.L., Wang, H., Hu, P., Yang, B., Pu, W., Shen, H., Chen, X., Liu, Z., Yin, H., Tan, Q., Wang, K., Hu, D., 2018. Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. EBioMedicine 30, 74–85.

Zhao, J., Huang, J., Zhi, D., Yan, W., Ma, X., Yang, X., Li, X., Ke, Q., Jiang, T., Calhoun, V. D., Sui, J., 2020. Functional network connectivity (FNC)-based generative adversarial network (GAN) and its applications in classification of mental disorders. J. Neurosci. Methods 341, 108756.

Zipursky, R.B., Menezes, N.M., Streiner, D.L., 2014. Risk of symptom recurrence with medication discontinuation in first-episode psychosis: a systematic review. Schizophr. Res. 152 (2–3), 408–414.