Chapter 8. Deep learning-based approaches in MRI: from discriminative to generative

Weizheng Yan

Lab of Neuroimaging, National Institute on Alcohol Abuse and Alcoholism

Deep learning (DL) has achieved remarkable success in natural language processing and natural image analysis, demonstrating its powerful capabilities. Nevertheless, its application in neuroimaging presents some unique challenges, including high dimensionality, low signal-to-noise ratio, small sample sizes, and limited ground truth. This chapter briefly surveys the deep learning models tailored for neuroimaging, especially fMRI. The chapter begins by presenting the advantages of DL in comparison to conventional machine learning and the basic DL architectures. Two categories of DL architectures, deep discriminative models and deep generative models are introduced respectively, with their progress in neuroimaging applications including classification, regression, biomarker discovery, subtype discovery, multi-site harmonization, and brain-wide dynamic modeling. Finally, we discuss priority areas for future studies of DL in neuroimaging research.

Key Words: deep learning, brain dynamic, fMRI, discriminative, generative, manifold, biomarker, subtype, multi-site, disentanglement

8.1. Why deep learning?

Natural images, as well as other real-world datasets, often span low-dimensional manifolds relative to their feature set [1]. Machine learning methods aim to discover meaningful patterns, such as feature subsets or feature representations, which can be further used for decisions. However, standard machine-learning (SML) algorithms (*e.g.*, linear regression) have limited capacity to process natural data in their raw form. Besides, constructing a machine learning system required careful engineering and considerable domain expertise to design feature extractors for converting the raw features (*e.g.,* pixel values of an image) into a suitable internal representation from which the learning subsystem, often a classifier, could recognize patterns[2-5].

Compared to SML, deep learning (DL) is a multi-layer representation learning method constructed by combining simple yet nonlinear units. Typically trained with error backpropagation, DL can approximate very complex mappings. DL has undergone significant developments in the past decade, attracting considerable interest from academia, industry, and funding agencies. Open source deep learning software frameworks (*e.g.,* Tensorflow https://www.tensorflow.org and Pytorch https://pytorch.org) and the explosion of available high performance computing infrastructure, especially graphics processing units also accelerated this development. With continuous emergence of state-of-the-art

network architectures and training techniques, neural networks can now be trained deeper [6] than ever before, resulting in improved performance. DL has outperformed SML in a wide range of applications such as computer vision [6], natural language processing [7], and speech recognition [8].

In the field of neuroimaging, deep learning has also exhibited its advantage in tasks such as classification of neuropsychiatric disorders, biomarker identification, and subtype discovery [9-11]. The availability of large imaging datasets, such as the Human Connectome Project, Adolescent Brain Cognitive Development Study, and UK Biobank, makes training DL models more viable. Neuroimaging samples differ from natural images in various ways (as shown in **Table 1**). For example, in contrast to natural images which are collected under natural light, neuroimaging data mostly consist of radiological images. Hence the noise distribution of neuroimaging varies depending on the acquisition used (e.g., Rician noise in MRI [12], quantum noise in computed tomography [13]). In addition, neuroimaging samples often exhibit multiple modalities, high dimensionality, low signal-to-noise ratio, and small sample sizes compared to natural images.

Table 1. Differences between natural images and neuroimaging samples.

	natural images	**neuroimaging samples**
data acquisition	Cheap for sample acquisition. Datasets can have over a million available samples.	Costly for sample acquisition. Datasets usually have fewer than 10^3 samples.
feature characteristics	Features are usually 3D images or videos under natural lighting; Noise is mostly Gaussian distributed.	Features are usually 3-D structures or four-dimensional time series. Mostly radiographic images. Noise distribution varies.
data annotation	Clear ground truth. Sample annotation is easy and generally does not require experts.	Usually lack clear ground truth. Annotation requires expert knowledge.
model training	Pre-trained models are usually available.	Few publicly available pre-trained models are available.
model interpretation	The effectiveness of the interpretable algorithm can be assessed intuitively.	The effectiveness of the interpretable algorithm is difficult to assess intuitively.

DL encompasses a collection of multi-layer architectures trained using error backpropagation approaches. Designing appropriate DL architectures suitable for the feature characteristics and sample size is critical. DL's versatility leads to numerous task-specific and data-oriented architectures, which can be challenging for beginners to navigate. Just as intricate LEGO worlds are constructed from basic elements, various DL architectures are built from fundamental modules. To build an intuitive and broad understanding of deep learning models, this section outlines the fundamental mechanisms of most basic DL models and guides their application in neuroimaging contexts, specifically fMRI. Sections 2 and 3 further categorize DL methods into two groups, deep discriminative models and deep generative models, for more in-depth comparison. The basic modules or models of the DL are as follows:

1. Multilayer neural networks

The multilayer fully connected neural networks, also named vanilla neural networks (vanilla NN), trained using the gradient backpropagation approach, are the simplest and

most illustrative DL models for replacing engineered features with trainable multilayers. The vanilla NN can transform the input space to make the class of data linearly separable. Theoretically, a vanilla NN can approximate any continuous function or mapping on compact subsets of R^n, given appropriate weights and activation functions. However, fully connected layers in vanilla NN may cause redundancy of trainable parameters and result in overfitting, even though the effects can be remediated by L1/L2 regularization and Dropout techniques. Vanilla NNs are usually applied for modeling low-dimensional and less redundant inputs such as FNC vectors [14]. In addition, due to its flexibility, vanilla NN is often used as 'bricks' for composing more complicated DL architectures (*e.g.*, autoencoders, and generative adversarial networks).

2. Convolutional neural network and graph convolutional network

The Convolutional Neural Network (CNN) is one of the most widely used DL architectures, applied in almost all computer vision tasks. It was designed to process data in the form of multiple arrays, such as a color image consisting of three 2D arrays containing pixel intensities in the three-color channels. In a classical CNN, two or three stages of convolution, non-linearity, and pooling are stacked, followed by more convolutional and fully connected layers [15]. The role of convolutional layers is to detect local conjunctions of features from the previous layer, while the pooling layers are to merge semantically similar features into one. Four key ideas are behind CNN that exploit the properties of natural signals: local connections, shared weights, pooling, and deep layers [2]. Given its ability to capture spatial information, CNN is well-suited for processing 2D images or 3D voxel-based [10] images. Besides pixel or voxel-based images, neuroimaging studies often use non-Euclidean graph structures to depict the relationships between nodes, such as functional connectivity. Graph Convolutional Network (GCN) is a type of neural network architecture designed to capture the graph structures and aggregate node information from the neighborhoods in a convolutional manner, with fewer learnable parameters. Therefore, GCNs are useful in medical or biochemical applications with graph-structured data [16].

3. Recurrent neural network

Recurrent neural network (RNN) models the generic dynamic system, $\dot{x}(t) = F(x(t), u(t))$. The state of the dynamic system $x(t)$ is updated by a vector-valued function F, which is non-linear and potentially complicated, accepts optional input $u(t)$. To implement this, the RNN processes an input sequence one element at a time, maintaining its hidden units as a 'state vector' that implicitly contains information about the history of all the previous elements in the sequence. The long short-term memory (LSTM) and gated recurrent unit (GRU) which is more simplified, are two practical variants of RNN designed for overcoming the vanishing gradient problem. RNN is suitable for modeling sequential inputs, such as fMRI time courses [17].

4. Generative adversarial network

The Generative adversarial network (GAN) was proposed for modeling complex distributions to generate realistic samples. A GAN comprises of two adversarial modules: a generator ***G*** and a discriminator ***D***. The ***G*** has no direct access to real data, the only way it learns is through its interaction with the ***D***. The ***D*** has access to both the synthetic samples

by ***G*** and real samples. Error signal to the ***D*** is provided through a simple ground truth of knowing whether the data came from the real data or ***G***. The same error signal, via ***D***, can be used to optimize ***G***, leading it toward synthesizing data of better quality [18]. GAN is not a specific model but a type of model framework. All the previously mentioned DL modules, such as vanilla NN, CNN, or RNN, can be used as ***D*** or ***G*** in GAN. Training generative models typically require a large number of samples and advanced model architectures to avoid posterior collapse [19], a situation where the model generates samples from only a small part of the latent space. Despite the challenge, the ability to generate high-quality new samples makes GAN well-suited for solving complex problems such as multi-site neuroimaging harmonization[20, 21] and subtype discovery [22].

5. Encoder-decoder

Encoder-decoder represents a category of generative framework composed of two core components: an encoder, which compresses the input features into a latent space representation, and a decoder, which reconstructs the latent space representation back into its original input feature space. Variational Autoencoder (VAE) is a typical encoder-decoder model designed for generative tasks. In VAE, the encoder transforms input data into a set of numbers that represent a probability distribution, and the decoder then randomly picks points from this distribution to generate diverse outputs, making VAEs effective in exploring and representing the underlying low-dimensional manifold of the data[23, 24]. Another typical model built on encoder-decoder architecture is the Transformer [25], which has revolutionized the natural language processing tasks. The Transformer utilizes self-attention mechanisms to capture dependencies between words in a sentence, enabling better handling of long-range dependencies. The Transformer has also been adapted for image processing. For example, the Vision Transformer (ViT), which applies the Transformer to image patches, has achieved state-of-the-art results on various imaging benchmarks [26].

To accommodate readers from various fields and due to space constraints, detailed explanation of DL concepts and their mathematical foundations is omitted. For a comprehensive understanding of DL models and the mathematics behind them, readers are recommended to consult the referenced DL books [3, 27].

8.2. Deep discriminative models: finding the boundaries.

Discriminative models learn to define the boundaries between different classes within the data. By learning the conditional probability distribution $P(Y|X)$ of output Y (*e.g.*, class labels) given input X (*e.g.,* image features), the discriminative models focus on mapping the observed features directly to target classes. Conventional discriminative machine learning models, such as logistic regression, support vector machine, and random forests, typically operate on manually selected or linear weighted combinations of features. However, deep discriminative models, usually constructed by CNN or RNN modules, use hierarchical feature learning strategies to capture complex relationships and map the features into low-dimensional task-specific manifolds or subspaces where the class boundaries are easier to recognize than in the original feature space. By analyzing the low-dimensional manifold, subtypes can be thereafter obtained. Besides, biomarkers can be discovered by post-hoc analysis of the trained models.

8.2.1. Classification and regression

Classification and regression are two widely studied discriminative tasks. The difference between classification and regression tasks is whether the target variable is discrete (classification) or continuous (regression). DL models typically use one-hot encoding for labeling categories in the output layer, which does not assume a natural order between labels, making DL models flexible to multi-class classification tasks. In comparison to natural images, which are 2D images, neuroimaging typically exhibits higher feature dimensionality (*e.g.,* fMRI consists of sequential volumes each containing over 10^6 voxels), smaller sample sizes (mostly fewer than 10^3 samples), multiple data modalities, and a lack of solid ground truth. Despite no clear guideline for choosing between standard machine learning or DL models, large sample sizes generally benefit DL models more than SMLs [10].

DL models should be adapted to the specific characteristics of the images to be studied. **Figure 1** summarizes the correspondence between neuroimaging features and DL models. Structural neuroimaging data reflect voxel tissue density (*e.g.,* T1-weighted) or structural connectivity (*e.g.,* diffusion MRI). One important research topic in structural studies is to establish relationships between structural features and symptoms, which can be used for clinical diagnosis or treatment response prediction. Given the structural MRI has 3D structural information, the 3D CNN model is the most intuitive choice. Abrol et al's study provides evidence that in age and gender classification tasks, with T1w samples over 300, the performance of DL has substantial improvement over SMLs [10].

In comparison to structural MRI, functional MRI has relatively lower spatial resolution, but includes temporal information, offering greater flexibility for analysis using DL. Even though the DL models have the advantage of extracting high-level feature representation from raw data, due to high dimensionality and low signal-noise-ratio in fMRI, efficient feature processing is still critical for reducing redundancy before modeling [28]. Specifically, fMRI raw data are often dimensionality-reduced using seed-based or data-driven approaches. The resulting temporal signatures are then used for studying temporal dependence such as functional network connectivity (FNC) or dynamic FNC. Different from natural images in which each voxel has conjunctions with it neighboring voxels, the FNC represents non-Euclidean graphical relationships. To analyze the FNC matrix as flattened 1D vector features, the vanilla NN is a viable solution. However, to analyze the FNC in its entirety while preserving its graph information, GCN offers a more suitable solution [29]. As for dynamic FNC, since it contains sequential information, RNN is more suitable for modeling. RNNs have achieved great successes in sequence for brain disorder diagnosis, brain decoding, and temporally dynamic functional state translation detection [30, 31]. Functional connectivity is often assessed using the Pearson correlation coefficient, which assumes a linear relationship between brain regions and oversimplifies their interactions. To address this limitation, a 1D convolutional module can be employed to automatically learn the non-linear relationship between regions, followed by an RNN to capture the temporal information for brain disorder classification [17].

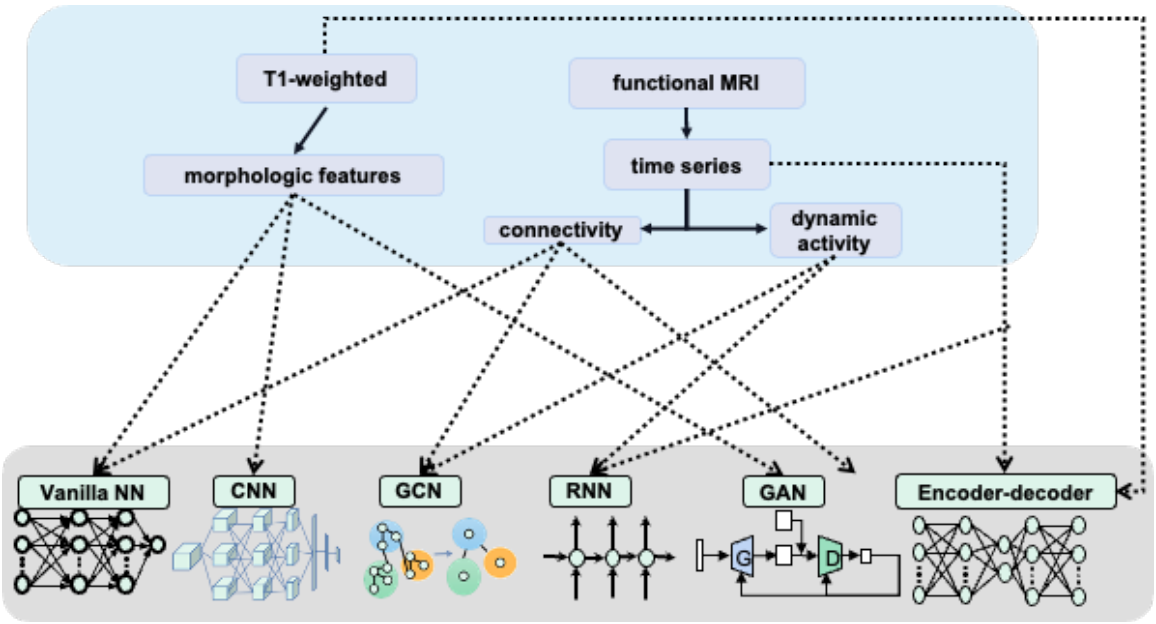

Figure 1. Different MRI features and their corresponding DL models. In the gray panel, multiple deep learning modules are listed and linked with their applicable features. *Abbreviations*: MLP: multi-layer perceptron; CNN: convolutional neural network; GCN: graph convolutional network; RNN: recurrent neural network. GAN: generative adversarial network.

8.2.2. Subtype discovery

Identifying disease or condition subtypes is crucial for improving our understanding of the underlying heterogeneity of neuropsychiatric disorders, personalized medicine, and targeted prevention interventions. To identify subtypes using neuroimaging, it is essential to detect patterns and correlations in brain structure and function that consistently correlate with specific variations in clinical symptoms and treatment response. The empirical success of the subtype identification is attributed to the manifold hypothesis: high-dimensional datasets are typically clustered near low-dimensional manifolds. In the case of fMRI analysis, due to its low signal-noise ratio and high dimensionality, confounding factors such as age, gender, or site effects may mislead clustering models. To address this, methods first need to eliminate irrelevant variables or confounding factors, and then apply clustering algorithms (*e.g.*, K-means, hierarchical clustering) to group samples based on similarity metrics. When employing standard machine learning for subtype discovery, the features should be carefully selected and purified. For example, when identifying major depressive disorder subtypes based on functional connectivity, canonical component analysis (CCA) could be employed to first map the functional connectivity to a syndrome-related subspace. The mapped functional connectivity features are then sent to a clustering model for further subtype discovery[5]. However, as a linear model, CCA may not sufficiently capture the full complexity of the data or include critical information necessary for subtype discovery. DL can extract features hierarchically at multiple levels of abstraction through its multilayer architecture and backpropagation training strategy. The architecture and training technique empower DL with a special characteristic: learning the continuous severity of conditions from binary-labeled images through its ability to identify intricate patterns and relationships among the samples. For example, as shown in **Figure**

2a, the T1w MRI dataset was divided into ten distinct groups based on gender and age range. One-hot encoding was utilized to label each sample and avoid any hierarchical ranking. A CNN model was then trained to classify the samples. After training, the features of the hidden layer for each sample were extracted and visualized using t-distributed stochastic neighbor embedding (t-SNE) on a 2D plane. The 2D projection spectrum was color-coded by the class labels, revealing separate gender clusters ordered in increasing age from one end of the spectrum to the other [10]. Similar results were obtained when using DL to discriminate Huntington's disease based on T1w MRI [32].

The continuous spectrum discovered by DL represents a low-dimensional manifold mapped from the original feature space to a task-specific subspace. For example, if a DL model with fMRI time courses as input features is trained to distinguish psychiatric disorders from healthy controls, the DL model will map fMRI time courses into a psychiatric-specific subspace, suppressing other r potential confounds such as age and gender. As shown in Figure 2b, the supervised multiple categorical classification model was first trained using a supervised way to map the original fMRI time series to a subspace where the differences between psychiatric disorders are more pronounced. The high-level representations of the original features are then submitted to a t-SNE clustering model to visualize the group differences between disorders, leading to the discovery of the schizoaffective disorder subtypes [33].

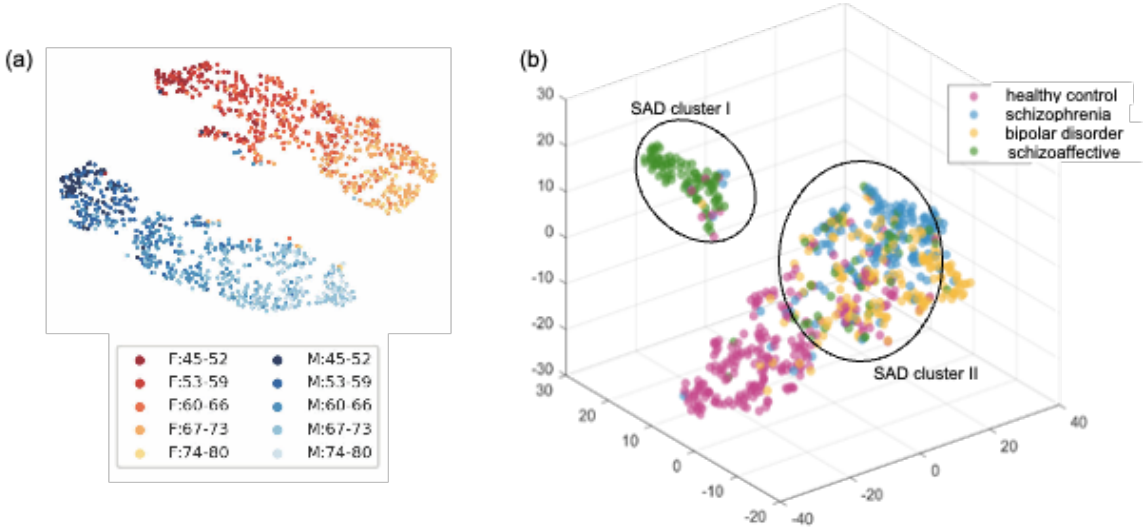

Figure 2. Subtype discovery using DL hidden layers. **(a)** The MRI embeddings inferred from a trained DL model used for age and gender classification. Representational patterns of the brain were learned. The samples eventually evolve into separate gender clusters (red/F/female and blue/M/male clusters), both presenting a gradual spectrum of age (traceable light-colored to dark-colored) [10]. **(b)** The fMRI embeddings inferred from a trained DL model used for multi-categorical mental disorder classification. The DL not only learned the relationships between mental disorders but also differentiated two schizoaffective subtypes.

8.2.3. Biomarker detection

The aim of identifying biomarkers is to establish relationships between easily understood features (*e.g.,* connectivity strength between two brain regions) and target

variables (*e.g.,* presence of schizophrenia). Biomarker discovery typically relies on the interpretation of DL models, a task that is challenging because the DL models use multiple non-linear layers to map features into subspaces. This complexity may lead to incorrect conclusions or interpretations, potentially limiting the clinical application of DL methods [30, 34].

For neuroimaging-based machine learning models to be interpretable, they should: (*i*) be comprehensible to humans, (*ii*) provide useful information about what mental or behavioral constructs are represented in particular brain pathways or regions, and (*iii*) demonstrate that they are based on relevant neurobiological signal, not artifacts or confounds [35]. The need to enable model interpretation has led to the development of various model introspection approaches, which can be broadly categorized into two groups: model-sensitive [36] and model-agnostic [37]. These approaches have their relative benefits and pitfalls in addressing the requirements of different applications [38].

The model-sensitive interpretation consists mainly of two types of approaches: gradient-based, and layer-wise relevance propagation [39]. Gradient-based methods can be computed using automatic differentiation and require no modification of the original DL model. Identification of discriminative brain regions in a classification of schizophrenia spectrum disorder vs. controls has been performed using a specific gradient-based implementation [40]. However, gradient-based methods are often computationally expensive, especially when making the integration procedure precise. Layer-wise relevance propagation utilizes the layered structure of the neural network and operates iteratively to produce an explanation. This analysis is performed at the level of individual input samples, allowing for analysis at multiple levels of data granularity, from the level of the group down to the level of single subjects, trails, and time points [41].

Model-agnostic interpretation often involves perturbation analysis, which repeatedly tests the effects on a DL's outputs when occluding patches or features from the inputs. It consists mainly of occlusion sensitivity, model-agnostic Explanation (LIME), and Shapley Additive exPlanations (SHAP). Specifically, occlusion analysis has been applied to CNN and RNN-based models for measuring the contribution of each brain region in classification tasks. For instance, a deep convolutional recurrent neural network was first trained to identify schizophrenia from healthy controls using fMRI time courses extracted from 50 brain regions. To identify the biomarkers related to schizophrenia, the brain regions were iteratively covered one by one to record and rank the resulting decrease in the model's performance. This process led to the discovery of the most contributing region, the stratum, for schizophrenia classification [17]. LIME produces explanations of a DL by approximating it locally with a simpler model (*e.g.*, a linear one) around the input sample being interpreted and then producing an intuitive summary of the simpler model that can be easily interpreted. SHAP computes Shapley values by considering all possible feature subsets and their contributions to the prediction. Lombardi *et al.* utilized SHAP and LIME respectively to determine the contribution of each brain morphological descriptor to the final predicted age of each subject. SHAP was reported to provide more reliable explanations for morphological aging mechanisms [42].

8.3. Deep generative models: disentangling the targeted variables

"*What I cannot create I do not understand.*" - Richard Feynman

Deep generative models are neural networks with multiple hidden layers designed to model complex, high-dimensional probability distributions. They can be used to generate new data samples that follow the same distributions as the training data. In contrast to discriminative models, which learn the conditional probability distribution $P(Y|X)$ of the target variable Y (*e.g.,* class labels) given the observable variable X (*e.g.,* image features), generative models are statistical models of the joint probability distribution $P(X, Y)$ of X and Y. When successfully trained, deep generative models can estimate the likelihood of each observation and generate new samples from the underlying distribution. Given that real-world datasets span low-dimensional manifolds relative to their feature set [1], many large-scale generative learning models are designed to map between complex datasets and simplified latent representations. Training generative models are usually computationally expensive but have significant advantages. The development of deep generative models has become one of the most researched fields in artificial intelligence in recent years, Recent advances in deep generative models have led to the production of photorealistic artwork (*e.g.,* DALL·E), precise protein structures (*e.g.,* AlphaFold), and natural-sounding conversational text (*e.g.,* ChatGPT) [19, 43].

Why are deep generative models important for neuroimaging? The complexity of the human brain arises from numerous genetic and environmental interactions. Brain images are entangled with confounds including demographic variables (*e.g.,* age, gender), and scanner effects (*e.g.,* magnetic strength, scanner manufacture). The effects of the confounds are often non-linear, despite decades of neuroimaging studies using linear assumptions for modeling them. Deep generative models have the capability to model the joint probability distribution $P(X, Y)$, which enables them to derive the conditional probability $P(X|Y)$ with ease. For instance, given the text label 'an apple', the system can generate a corresponding image. A unique feature of deep generative models is variable disentanglement, which enables the extraction of specific attributes, such as site effects. This capability provides a powerful tool for isolating and examining individual factors, making it essential for tasks such as multi-site data harmonization or decoding brain activity [44]. The literatures [18, 19, 43, 45] extensively discussed primary deep generative models, including GAN, and encoder-decoder models, represented by VAE. GANs introduce a dynamic competition between two neural networks: a generator that produces synthetic data and a discriminator that evaluates the authenticity of both real and synthetic data, resulting in the creation of highly realistic samples. This adversarial process has shown significant success in various domains, such as image and voice generation. VAEs, on the other hand, are notable for their ability to create new samples that closely resemble the training data by using spherical Gaussian distributions in latent space. VAE guarantees that the acquired representations are proficiently encoded in terms of latent variables, revealing the complex underlying factors of the dataset. In this section, three applications of deep generative models in neuroimaging are introduced: subtype discovery, multi-site harmonization, and brain-wide dynamic modeling.

8.3.1. Subtype discovery

Brain development results from dynamic interactions between genes and environment, and brain imaging datasets are further complicated by factors such as scanner effects, age,

and gender. Variables unrelated to diseases may mislead clustering algorithms. Therefore, subtype discovery algorithms must be designed to avoid the influence of these disease-irrelevant variables. The previous section discussed discriminative model approaches that involves mapping the original features to subspaces related to the target variables, thereby identifying a low-dimensional disease-specific manifold. Clustering algorithms are then used for subtype discovery, significantly reducing the impact of irrelevant variables. However, discriminative models are not able to disentangle disease-relevant variables from irrelevant ones. To address this limitation, we mainly introduce two deep generative models, contrastive VAE (CVAE) [24] and SeMI-supervised cLustEring-Generative Adversarial Network (Smile-GAN) [22], for identifying subtypes based on structural MRI.

CVAE is an autoencoder model designed specifically for extracting latent variables for clustering purposes [24]. As shown in Fig. 3a, CVAE takes input samples from two different populations (healthy controls and mental disorders) and isolate variation specific to one population from variation common to both. This enables the CVAE to separate 'disorder-specific' neuroanatomical variation from variation shared by both mental disorder and healthy control, representing each as a distinct set of latent features. In comparison to conventional disentangling approaches and architectures, the CVAE has three significant advantages. First, the CVAE allows for modeling nonlinear relationships between inputs and latent features, providing more flexibility compared to linear methods such as contrastive PCA. Second, CVAE can separate shared and specific features in their latent space, even when the features are entangled in the inputs, setting them apart from some multimodal methods that rely on inputs where different modalities are already separated. Third, CVAE includes a decoder, enabling the implementation of the "synthetic twin" analysis. This decoder allows for generating counterfactual brains from the latent features, which can then be used to identify interpretable brain regions affected by structural variation within the patients.

Different from CVAE which separates generative and clustering stages, Smile-GAN integrates a generative module with a clustering module, addressing the need for information flow between these two components. As shown in Fig 3b, Smile-GAN uses GAN to learn the non-linear mappings from normal control to patients, allowing for a disentangle of disease-relevant and disease-irrelevant variables. Additionally, a clustering module is integrated into the training framework, enabling the model to group similar data points based on the learned features. This simultaneous training of the generator, discriminator, and clustering components ensures that the model not only generates realistic data but also effectively clusters it, enriching the overall analysis process [22].

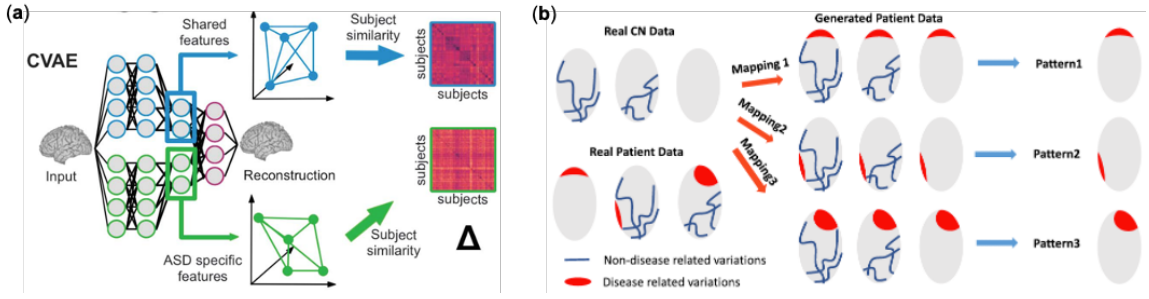

Figure3. (**a**) Architecture of the CVAE model. CVAE takes input samples from two distinct populations and isolates variation specific to one population from variation

common to both. Hence, CVAEs disentangle 'autism-specific' neuroanatomical variation from variation 'shared' by both autism and typical control participants, representing each as a distinct set of latent features. (**b**) A conceptual overview of Smile-GAN. Blue lines represent non-disease-related variations observed in both normal control and patient groups. Red regions represent disease effects that only exist in patient groups. Smile-GAN discovers disease-specific neuroanatomical patterns by learning non-linear transformations from normal controls to various diseases.

8.3.2. Multi-site datasets harmonization

Multi-site neuroimaging collaboration is a viable way to overcome small sample bias by aggregating samples from multiple sites or hospitals. However, samples from different sites are typically acquired using various scanners, acquisition protocols, and software versions. This variability partly explains the significant degradation of pooled classification performance as sample size increases. In addition, DL models have the ability to detect and leverage site-specific information for classification purposes [46], which may result in classifiers that are not generalizable or robust. Therefore, proper harmonization of site and scanner effects is critical to mitigate these differences for downstream analysis [47].

The key to solving the multi-site harmonization model is to accurately estimate the distribution of site-specific variables or to disentangle the site-specific variables from other image features. GANs are often employed to address multi-site harmonization challenges [20, 21, 48]. For instance, to harmonize T1w MRI samples collected from six sites, a GAN variant called StarGAN was employed. The StarGAN consists of a style encoder, a content encoder, a generator, and a discriminator. Both the content encoder and the generator are CNNs that map a single axial slice from a T1w scan to low-dimensional representations. Specifically, the style encoder learns a mapping of slices to eight-dimensional vectors representing the site-based variation of that slice. The content encoder learns a mapping of slices to a lower-dimensional set of convolutional filters representing the site-irrelevant information of the slice. The generator then combines both the style and content encodings to produce a harmonized image that maintains the respective style and content of the input encodings. This process generates a harmonized scan that matches the site-based variation of a reference scan while preserving the original anatomical information. The discriminator network is used to make the generated images more realistic [21].

It is worth noting that in the process of harmonizing site-specific variables, information relevant to downstream tasks may also be lost. Therefore, it is essential to consider the requirements of downstream tasks when designing DL multi-site harmonization methods [48].

8.3.3. Brain-wide dynamic modeling

A major tenet of theoretical neuroscience is that cognitive and behavioral processes are ultimately implemented through neural system dynamics. In many brain regions, the activity of a large population of neurons is often well described by low-dimensional dynamics. Advanced neural technologies allow for recording from many thousands of neurons in multiple interacting brain areas, enabling the manipulation and modeling of brain-wide neural population dynamics. Recovering these dynamics on single trials is essential for illuminating the relationship between neural population activity and behavior,

and for advancing therapeutic neurotechnology such as closed-loop deep brain stimulation and brain-machine interfaces. By combining deep learning with dynamic systems, the performance of neural systems can be significantly improved. Here, two instances are primarily introduced: Latent Factor Analysis via Dynamical Systems (LFADS)[49] which models neural spiking, and piecewise-linear recurrent neural network (PLRNN) [50] which models fMRI.

Dynamic system modeling aims to model the following generic dynamic system:

$$\dot{x}(t) = \mathrm{F}(\mathrm{x}(t), \mathrm{u}(t))$$

The state of the dynamical system, *x(t)*, is updated by the vector-valued function *F*, which is potentially complex, accepts option input *u(t)*, and is initialized by an initial condition *x(0)*. Traditional machine learning approaches for modeling neural population dynamics, such as linear dynamical systems (LDS), typically make simplifying assumptions by modeling the underlying population dynamics as independent, linear or switched linear. However, uncovering relevant transient patterns in brain function is challenging due to the lack of computational tools that can effectively capture nonlinear dynamics from high-dimensional data. Recent studies show that DL models, especially those based on RNN, have the potential to capture whole-brain dynamic information and exploit time-varying functional connectivity state profiles, advancing our understanding of brain function and diseases [51, 52].

The LFADS model is a deep generative model designed to infer underlying dynamics from single-trial neural spiking data. It operates on the assumption that observed neural activity arises from a lower-dimensional dynamical system characterized by latent factors. As shown in Figure 4a, LFADS employs a sequential autoencoder architecture consisting of a variational autoencoder extended to sequences. This architecture includes an encoder (to compress observed data into latent representations), a generator (to generate dynamics and inferred firing rates from these representations), and optionally a controller (to model external inputs to the neural population). Through this approach, LFADS is able to de-noise spiking activity, predict behavioral variables, and extract precise estimates of single-trial neural dynamics, thereby providing insight into the neural computation underlying observed behaviors [49].

Deep generative models can also advance the identification of the computational dynamics underlying task processing [45]. Koppe et al [50] proposed an advanced state space model (SSM) based on generative piecewise-linear recurrent neural networks (PLRNNs) for analyzing neuroimaging data, specifically fMRI. The PLRNN forces the latent model to capture the 'true' underlying dynamics rather than just fitting (or predicting) the observations. As shown in Figures 4b and 4c, the approach demonstrated the ability to uncover task-related nonlinear structures that linear models fail to capture, providing a novel step towards analyzing non-linear dynamics in neuroscientific research and clinical

assessment. [50]

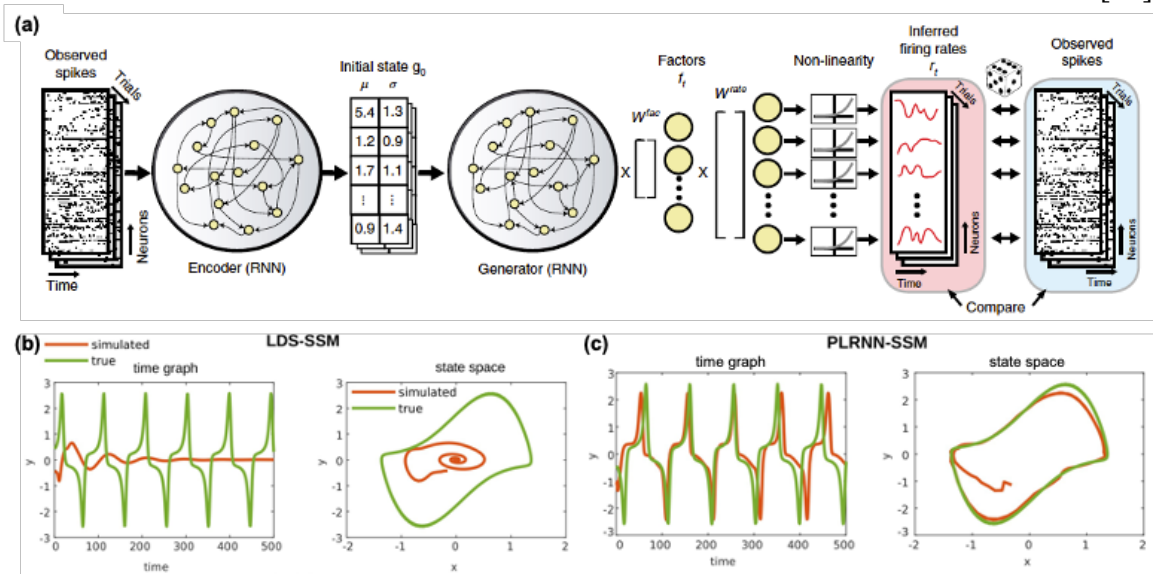

Figure 4. (**a**) schematic overview of the LFADS architecture[49]. (**b**) Example time series from an LDS-SSM and a PLRNN-SSM trained on the van der Pol (vdP) system. Example time graph (left) and state space (right) for a trajectory generated by an LDS-SSM (red) trained on the vdP system (true vdP trajectories are in green). Trajectories from an LDS will almost inevitably decay toward a fixed point over time (or diverge). (**c**) Trajectories generated by a trained PLRNN-SSM, in contrast, closely follow the vdP-system's original limit cycle [50].

8. 4. Further study

8.4.1. The balance between task complexity and model complexity

DL models are growing in number and complexity as the field of neuroimaging advances. They are now being applied to different levels of features and a range of tasks in neuroimaging. It is commonly observed that the use of more primitive features and the performance of more complex tasks require increasingly sophisticated algorithms and correspondingly larger training datasets. For example, DL for gender classification based on functional connectivity features is generally simpler and requires less data than 3D CNN-LSTM algorithms trained on raw fMRI data for identifying dynamic attractors. Although some auto-differentiation platforms (*e.g.,* Pytorch, TensorFlow) have greatly simplified model design procedures, various hyper-parameters such as width, depth, loss function, and optimizers are typically decided based on experience.

This empirical knowledge remains unquantified, presenting a gap between practical and theoretical understanding. In addition, different tasks require varying levels of performance from these models. For example, a DL model designed for cancer screening usually requires better performance, especially sensitivity, than a DL model designed for influenza screening. In the future, there is a need for a theoretical framework capable of determining the minimal complexity of models based on performance requirements and training samples. Such a framework would also need to establish the lower bound for the amount of training data required. This framework would not only streamline the development of efficient models tailored to specific neuroimaging tasks but would also

provide a quantitative basis for the empirical insights gained from the use of DL in neuroimaging.

8.4.2. Large language model (LLM) and biological mechanisms

Large Language Models (LLMs), such as ChatGPT, built on large neural networks and trained on extensive corpora of text, have demonstrated remarkable text processing capabilities and are now the best-performing model for automatic translation, summarization, dialogue, and even complex reasoning tasks. Despite their success, the performance of LLMs often relies on statistical patterns and associations learned from the training data, rather than an understanding of underlying principles or meanings. The working mechanisms of LLMs, particularly how they manage and represent information and how similar they are to human cognition, represent a significant research challenge. By conducting comparative studies between LLMs and neuroimaging, we can effectively bridge biological mechanisms with algorithms. This comparison between computational models and the biological brain allows for a deeper understanding of the underlying computations that drive brain function [53-55]. Aligning LLMs with biological mechanisms may also allow the development of interpretable algorithms for biological data that go beyond the 'black box' models of current DL, leading to superior explanatory power in medical and health-related research.

8.4.3. Neuromodulatory systems and multi-level analysis

To achieve a comprehensive understanding of brain dynamics across different levels, from molecular to behavioral, it is crucial to establish multilayered computational models [56, 57]. The study of neuromodulatory systems has shown their potential to bridge the gap between structure and function [58]. By connecting different levels of description, computational models can parametrically map classical neuromodulatory processes onto system-level models of neural activity, enriching our understanding beyond what can be achieved through structure or functional connectivity alone. Moreover, the success of DL and the similarities in their structural organization to the cortex suggest that humans and deep machines may share fundamental computational principles [57]. Incorporating DL models can future improve the performance of current computational models. This interdisciplinary approach not only paves the way for DL models that more accurately mimic the brain's inherent mechanisms but also enhances our ability to model and understand the intricate dynamics of neuromodulation in shaping neural and cognitive functions.

Acknowledgements

This research was supported by the Intramural Research Program of the NIH. The authors thanks Dr. Nora Volkow for the valuable comments.

References

[1] C. Fefferman, S. Mitter, and H. Narayanan, "Testing the manifold hypothesis," *Journal of the American Mathematical Society,* vol. 29, no. 4, pp. 983-1049, 2016.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436-44, May 28 2015, doi: 10.1038/nature14539.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[4] M. Xu, V. Calhoun, R. Jiang, W. Yan, and J. Sui, "Brain imaging-based machine learning in autism spectrum disorder: methods and applications," *Journal of Neuroscience Methods,* vol. 361, p. 109271, 2021/09/01/ 2021, doi: https://doi.org/10.1016/j.jneumeth.2021.109271.

[5] A. T. Drysdale *et al.*, "Resting-state connectivity biomarkers define neurophysiological subtypes of depression," *Nature medicine,* vol. 23, no. 1, pp. 28-38, Jan 2017, doi: 10.1038/nm.4246.

[6] X. Z. Kaiming He, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition," presented at the CVPR, 2016.

[7] T. Wu *et al.*, "A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development," *IEEE/CAA Journal of Automatica Sinica,* vol. 10, no. 5, pp. 1122-1136, 2023, doi: 10.1109/jas.2023.123618.

[8] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimedia Tools and Applications,* vol. 80, no. 6, pp. 9411-9457, 2021/03/01 2021, doi: 10.1007/s11042-020-10073-7.

[9] W. Yan *et al.*, "Deep Learning in Neuroimaging: Promises and challenges," *IEEE Signal Processing Magazine,* vol. 39, no. 2, pp. 87-98, 2022, doi: 10.1109/MSP.2021.3128348.

[10] A. Abrol *et al.*, "Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning," *Nature Communications,* vol. 12, no. 1, p. 353, 2021/01/13 2021, doi: 10.1038/s41467-020-20655-6.

[11] S. M. Plis *et al.*, "Deep learning for neuroimaging: a validation study," (in English), *Front Neurosci,* Methods vol. 8, 2014-August-20 2014, doi: 10.3389/fnins.2014.00229.

[12] I. I. Maximov, E. Farrher, F. Grinberg, and N. Jon Shah, "Spatially variable Rician noise in magnetic resonance imaging," *Medical Image Analysis,* vol. 16, no. 2, pp. 536-548, 2012/02/01/ 2012, doi: https://doi.org/10.1016/j.media.2011.12.002.

[13] A. J. Duerinckx and A. Macovski, "Information and artifact in computed tomography image statistics," *Med Phys,* vol. 7, no. 2, pp. 127-134, 1980/03/01 1980, doi: https://doi.org/10.1118/1.594771.

[14] J. Kim, V. D. Calhoun, E. Shim, and J. H. Lee, "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,* vol. 124, no. Pt A, pp. 127-46, Jan 01 2015, doi: 10.1016/j.neuroimage.2015.05.018.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* vol. 25, 2012.
[16] M. Haghir Chehreghani, "Half a decade of graph convolutional networks," *Nature Machine Intelligence,* vol. 4, no. 3, pp. 192-193, 2022/03/01 2022, doi: 10.1038/s42256-022-00466-8.
[17] W. Yan *et al.*, "Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site FMRI data," *EBioMedicine,* vol. 47, pp. 543-552, Sep 2019, doi: 10.1016/j.ebiom.2019.08.023.
[18] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative Adversarial Networks: An Overview," *IEEE Signal Processing Magazine,* vol. 35, no. 1, pp. 53-65, 2018, doi: 10.1109/msp.2017.2765202.
[19] W. Gilpin, "Generative learning for nonlinear dynamics," *Nature Reviews Physics,* 2024, doi: 10.1038/s42254-024-00688-2.
[20] S. Liu and P.-T. Yap, "Learning multi-site harmonization of magnetic resonance images without traveling human phantoms," *Communications Engineering,* vol. 3, no. 1, p. 6, 2024/01/05 2024, doi: 10.1038/s44172-023-00140-w.
[21] V. M. Bashyam *et al.*, "Deep Generative Medical Image Harmonization for Improving Cross-Site Generalization in Deep Learning Predictors," *Journal of Magnetic Resonance Imaging,* vol. n/a, no. n/a, doi: https://doi.org/10.1002/jmri.27908.
[22] Z. Yang *et al.*, "A deep learning framework identifies dimensional representations of Alzheimer's Disease from brain structure," *Nat Commun,* vol. 12, no. 1, p. 7065, Dec 3 2021, doi: 10.1038/s41467-021-26703-z.
[23] J.-H. Kim, Y. Zhang, K. Han, Z. Wen, M. Choi, and Z. Liu, "Representation learning of resting state fMRI with variational autoencoder," *NeuroImage,* vol. 241, p. 118423, 2021/11/01/ 2021, doi: https://doi.org/10.1016/j.neuroimage.2021.118423.
[24] A. Aglinskas, J. K. Hartshorne, and S. Anzellotti, "Contrastive machine learning reveals the structure of neuroanatomical variation within autism," *Science,* vol. 376, no. 6597, pp. 1070-1074, 2022, doi: doi:10.1126/science.abm2461.
[25] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.
[26] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929,* 2020.
[27] C. M. Bishop and H. Bishop, *Deep learning: Foundations and concepts*. Springer Nature, 2023.
[28] W. Yin, L. Li, and F.-X. Wu, "Deep learning for brain disorder diagnosis based on fMRI images," *Neurocomputing,* vol. 469, pp. 332-345, 2022/01/16/ 2022, doi: https://doi.org/10.1016/j.neucom.2020.05.113.
[29] A. Bessadok, M. A. Mahjoub, and I. Rekik, "Graph Neural Networks in Network Neuroscience," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, no. 5, pp. 5833-5848, 2023, doi: 10.1109/TPAMI.2022.3209686.

[30] M. M. Rahman *et al.*, "Interpreting models interpreting brain dynamics," *Scientific Reports,* vol. 12, no. 1, p. 12023, 2022/07/21 2022, doi: 10.1038/s41598-022-15539-2.
[31] W. Yan, H. Zhang, J. Sui, and D. Shen, "Deep Chronnectome Learning via Full Bidirectional Long Short-Term Memory Networks for MCI Diagnosis," *Med Image Comput Comput Assist Interv,* vol. 11072, pp. 249-257, Sep 2018, doi: 10.1007/978-3-030-00931-1_29.
[32] S. M. Plis *et al.*, "Deep learning for neuroimaging: a validation study," *Front Neurosci,* vol. 8, p. 229, 2014, doi: 10.3389/fnins.2014.00229.
[33] W. Yan, M. Zhao, Z. Fu, G. D. Pearlson, J. Sui, and V. D. Calhoun, "Mapping relationships among schizophrenia, bipolar and schizoaffective disorders: A deep classification and clustering framework using fMRI time series," *Schizophrenia Research,* vol. 245, pp. 141-150, 2022/07/01/ 2022, doi: https://doi.org/10.1016/j.schres.2021.02.007.
[34] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.
[35] L. Kohoutová *et al.*, "Toward a unified framework for interpreting machine-learning models in neuroimaging," *Nature Protocols,* vol. 15, no. 4, pp. 1399-1435, 2020/04/01 2020, doi: 10.1038/s41596-019-0289-5.
[36] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825,* 2017.
[37] M. T. Ribeiro, S. Singh, and C. Guestrin, "" Why should i trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135-1144.
[38] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proceedings of the IEEE,* vol. 109, no. 3, pp. 247-278, 2021.
[39] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K. R. Müller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proceedings of the IEEE,* vol. 109, no. 3, pp. 247-278, 2021, doi: 10.1109/JPROC.2021.3060483.
[40] K. Oh *et al.*, "Classification of schizophrenia and normal controls using 3D convolutional neural network and outcome visualization," *Schizophrenia Research,* vol. 212, pp. 186-195, 2019/10/01/ 2019, doi: https://doi.org/10.1016/j.schres.2019.07.034.
[41] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Muller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLoS One,* vol. 10, no. 7, p. e0130140, 2015, doi: 10.1371/journal.pone.0130140.
[42] A. Lombardi *et al.*, "Explainable Deep Learning for Personalized Age Prediction With Brain Morphology," (in English), *Front Neurosci,* Original Research vol. 15, no. 578, 2021-May-28 2021, doi: 10.3389/fnins.2021.674055.
[43] L. Ruthotto and E. Haber, "An introduction to deep generative modeling," *GAMM-Mitteilungen,* vol. 44, no. 2, 2021, doi: 10.1002/gamm.202100008.

[44] R. VanRullen and L. Reddy, "Reconstructing faces from fMRI patterns using deep generative neural networks," *Communications Biology,* vol. 2, no. 1, p. 193, 2019/05/21 2019, doi: 10.1038/s42003-019-0438-y.
[45] M. Ramezanian-Panahi, G. Abrevaya, J.-C. Gagnon-Audet, V. Voleti, I. Rish, and G. Dumas, "Generative Models of Brain Dynamics," (in English), *Frontiers in Artificial Intelligence,* Review vol. 5, 2022-July-15 2022, doi: 10.3389/frai.2022.807406.
[46] Z. Liu and K. He, "A Decade's Battle on Dataset Bias: Are We There Yet?," *arXiv preprint arXiv:2403.08632,* 2024.
[47] H. Guan and M. Liu, "Domain Adaptation for Medical Image Analysis: A Survey," *IEEE Trans Biomed Eng,* vol. 69, no. 3, pp. 1173-1185, Mar 2022, doi: 10.1109/TBME.2021.3117407.
[48] W. Yan, Z. Fu, R. Jiang, J. Sui, and V. D. Calhoun, "Maximum Classifier Discrepancy Generative Adversarial Network for Jointly Harmonizing Scanner Effects and Improving Reproducibility of Downstream Tasks," *IEEE Transactions on Biomedical Engineering,* pp. 1-9, 2023, doi: 10.1109/TBME.2023.3330087.
[49] C. Pandarinath *et al.*, "Inferring single-trial neural population dynamics using sequential auto-encoders," *Nature Methods,* vol. 15, no. 10, pp. 805-815, 2018/10/01 2018, doi: 10.1038/s41592-018-0109-9.
[50] G. Koppe, H. Toutounji, P. Kirsch, S. Lis, and D. Durstewitz, "Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fMRI," *PLOS Computational Biology,* vol. 15, no. 8, p. e1007263, 2019, doi: 10.1371/journal.pcbi.1007263.
[51] U. Mahmood, Z. Fu, V. D. Calhoun, and S. Plis, "A Deep Learning Model for Data-Driven Discovery of Functional Connectivity," *Algorithms,* vol. 14, no. 3, p. 75, 2021.
[52] B. Kazemivash and V. D. Calhoun, "A novel 5D brain parcellation approach based on spatio-temporal encoding of resting fMRI data from deep residual learning," *bioRxiv,* p. 2021.04.22.440936, 2021, doi: 10.1101/2021.04.22.440936.
[53] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature,* vol. 620, no. 7972, pp. 172-180, 2023/08/01 2023, doi: 10.1038/s41586-023-06291-2.
[54] A. G. Russo, A. Ciarlo, S. Ponticorvo, F. Di Salle, G. Tedeschi, and F. Esposito, "Explaining neural activity in human listeners with deep learning via natural language processing of narrative text," *Scientific Reports,* vol. 12, no. 1, p. 17838, 2022/10/25 2022, doi: 10.1038/s41598-022-21782-4.
[55] C. Caucheteux, A. Gramfort, and J.-R. King, "Deep language algorithms predict semantic comprehension from brain activity," *Scientific Reports,* vol. 12, no. 1, p. 16327, 2022/09/29 2022, doi: 10.1038/s41598-022-20460-9.
[56] J. M. Shine *et al.*, "Human cognition involves the dynamic integration of neural activity and neuromodulatory systems," *Nature Neuroscience,* vol. 22, no. 2, pp. 289-296, 2019/02/01 2019, doi: 10.1038/s41593-018-0312-0.
[57] J. M. Shine, E. J. Müller, B. Munn, J. Cabral, R. J. Moran, and M. Breakspear, "Computational models link cellular mechanisms of neuromodulation to large-

scale neural dynamics," *Nature Neuroscience,* vol. 24, no. 6, pp. 765-776, 2021/06/01 2021, doi: 10.1038/s41593-021-00824-6.

[58] M. L. Kringelbach *et al.*, "Dynamic coupling of whole-brain neuronal and neurotransmitter systems," *Proc Natl Acad Sci U S A,* vol. 117, no. 17, pp. 9566-9576, Apr 28 2020, doi: 10.1073/pnas.1921475117.