Chapter 8. Deep Learning for Brain Dynamics: from Discriminative to Generative Models

Weizheng Yan[1,2]

[1]*Lab of Neuroimaging, National Institute on Alcohol Abuse and Alcoholism*
[2]*Department of Neurology, Renaissance School of Medicine, Stony Brook, NY*

Deep Learning (DL) provides powerful tools for modeling complex spatiotemporal patterns in neuroimaging data, offering new opportunities for understanding brain activities. However, applying DL to this field remains challenging because neuroimaging is often high dimensional, noisy, limited in sample size, and affected by substantial inter-subject variability. This chapter reviews major DL models for brain dynamics analysis. It first introduces fundamental DL modules and then organizes existing DLs into two broad categories: deep discriminative models and deep generative models. The chapter then reviews DL applications across key topics, including disease classification, and prediction, biomarker discovery, subtype identification, multi-site harmonization, and latent brain state analysis. Finally, it discusses future directions including balancing task and model complexity, incorporating neuromodulation-informed multi-level analysis, and developing foundation models for brain dynamics.

Key Words: deep learning, brain dynamics, discriminative, generative, biomarker, subtype, data harmonization

8.1. Overview of deep learning in neuroimaging

Real-world data, such as natural images, often lie on low-dimensional manifolds embedded in high-dimensional feature spaces. Machine learning methods aim to uncover meaningful patterns in data, such as informative feature subsets or latent feature embeddings, which can subsequently support decision-making. Standard machine-learning (SML) algorithms (*e.g.,* nonlinear regression) often have limited ability to model complex patterns of high-dimensional data. Building SML systems typically requires careful engineering and domain expertise to design feature extractors that transformed the raw feature (*e.g.,* image pixel) into suitable feature representations from which a learning subsystem can further identify patterns (Drysdale et al., 2017; Goodfellow et al., 2016; LeCun et al., 2015; Xu et al., 2021). In contrast to SML, deep learning (DL) learns hierarchical representations by applying multiple layers of nonlinear transformations. These layers progressively map raw input features into increasingly abstract representations, allowing the model to capture complex patterns in the data. Through gradient-based backpropagation, DL models optimize these transformations jointly and can approximate highly complex input-output relationships. Advances in high-performance computing, open-sourced DL platforms (*e.g.,* Tensorflow and Pytorch), and training techniques have enabled the successful training of efficient DL models, leading

to improved performance. DL models have outperformed SML in a wide range of applications such as image classification (He et al., 2016), natural language processing (Wu et al., 2023), and speech recognition (Malik et al., 2021). MRI studies also indicate that DL can learn better discriminative feature representations than SML (Abrol et al., 2021).

In the field of neuroimaging analysis, DL has also exhibited its advantage in tasks such as disease classification, biomarker identification, and subtype discovery (Abrol et al., 2021; Han et al., 2026; Rahman et al., 2026; Thapaliya et al., 2025; Yan et al., 2022). The availability of large-scale neuroimaging datasets (*e.g.,* Human Connectome Project, Adolescent Brain Cognitive Development Study, and UK Biobank) has made training large-scale DL models feasible. As shown in Table 1, neuroimaging data differs from natural images in several aspects. For example, unlike natural images, which are captured under natural lighting conditions, neuroimaging data are acquired using specialized radiological modalities, each with distinct physical principles and noise characteristics. For example, MRI data may be affected by Rician noise, whereas CT images are commonly influenced by quantum noise. In addition, neuroimaging datasets often have high dimensionality, relatively low signal-to-noise ratios, and much smaller sample sizes than large-scale natural image datasets. Together, these characteristics pose major challenges for developing and applying DL in neuroimaging (Yan et al., 2022).

Table 1. Differences between natural image datasets and neuroimaging datasets.

	Natural image datasets	Neuroimaging datasets
Data acquisition	Low acquisition cost. Large datasets with millions of samples are common.	High acquisition cost due to expensive imaging equipment and clinical procedures. Datasets typically contain fewer than 10^4 samples.
Data characteristics	Images captured under natural lighting conditions; Noise is mostly Gaussian distributed.	Radiographic images acquired using specialized scanners (*e.g.,* MRI, CT, PET). Often 3D volumes or 4D time series.
Feature properties	Visual features are relatively stable and directly observable	Features are often subtle and influenced by physiological and acquisition-related factors
Noise properties	Often approximate as Gaussian	Depending on image modality (*e.g.,* Rician in MRI, quantum in CT)
Data annotation	Clear ground truth labels are available. Annotation is relatively easy and typically does not require expert knowledge.	Ground truth is often ambiguous or unavailable. Annotation typically requires expert knowledge.
Model training	Pre-trained models are widely available.	Few publicly available pre-trained models exist.
Model interpretation	Model interpretation can often be assessed intuitively by visual inspection.	Model interpretation is more challenging due to complex anatomical and physiological factors.

DL refers to a family of methods based on multilayer neural network architectures optimized through gradient-based backpropagation. Because neuroimaging data have distinct characteristics, such as high dimensionality, spatial structure, temporal dependency, and limited sample sizes, DL models must be carefully designed to match both the properties of the training data and the goals of specific neuroimaging tasks. Similar to LEGO structures built from basic components, DL architectures are constructed from fundamental modules that can be combined in different ways for different applications. To facilitate understanding, particularly for readers new to DL, this

section introduces the core mechanisms of commonly used DL modules and discusses their relevance to neuroimaging applications, with a focus on MRI. Specifically, we review fundamental architectures, including multilayer neural networks, convolutional neural networks, graph convolutional networks, recurrent neural networks, and encoder–decoder models, as well as broader generative frameworks, including generative adversarial networks and diffusion models.

1. Multilayer neural networks

Multilayer fully connected neural networks, also known as vanilla neural networks (vanilla NNs), are among the simplest DL models. Trained through gradient-based backpropagation, they learn nonlinear transformations and, in theory, can approximate arbitrary functions. However, because each neuron is connected to all neurons in adjacent layers, fully connected architectures often contain a large number of trainable parameters. This can introduce redundancy and increase the risk of overfitting, especially when applied to high-dimensional neuroimaging data with limited sample sizes. These limitations can be partially mitigated using techniques such as L1/L2 regularization and dropout. Despite these challenges, vanilla NNs are useful for modeling relatively low-dimensional and less redundant features, such as functional connectivity (FC) (Kim et al., 2015). Owing to their flexibility, vanilla NNs are also widely used as building blocks in more complex deep learning architectures, including autoencoders and generative adversarial networks.

2. Convolutional neural networks (CNNs) and graph neural networks (GNNs)

CNNs are among the most widely used DL models and have been extensively applied to image processing tasks. CNNs are designed to process data organized as arrays, such as color images such as color images represented by multiple two-dimensional channels or three-dimensional neuroimaging volumes. A typical CNN consists of stacked convolutional layers, nonlinear activation functions, and pooling layers, followed by fully connected layers for prediction. Convolutional layers extract local feature patterns from local receptive fields, whereas pooling layers aggregate semantically similar features and progressively enlarge the effective receptive field. Through this hierarchical structure, CNNs integrate local features into increasingly global and abstract representations, enabling them to model both fine-grained and large-scale spatial structures. Because of this ability to leverage spatial organization, CNNs are well suited for processing two-dimensional brain images and three-dimensional neuroimaging volumes (Abrol et al., 2021). In addition to image- or volume-based representations, neuroimaging studies often use non-Euclidean graph structures, such as FC, to characterize brain organization. GNNs are DL models specifically designed for such graph-structured data. Following a message-passing framework, GNNs update node representations by aggregating information from neighbouring nodes, thereby capturing relational patterns within the graph (Tang et al., 2026). This makes GNNs particularly suitable for neuroimaging applications involving graph-based brain representations, including FC analysis (Han et al., 2026).

3. Recurrent neural networks (RNNs)

Recurrent neural networks (RNNs) are designed to model sequential data by updating an internal hidden state over time. Conceptually, they can be viewed as a generic dynamical system of the form $\dot{x}(t) = F(x(t), u(t))$, where the system state $x(t)$ evolves according to a vector-valued nonlinear function F, potentially influenced by an external input $u(t)$. In practical applications, RNNs process input sequences one element at a time, while their hidden states serve as state vectors that implicitly encode information from previous time points. However, traditional RNNs are limited by the vanishing gradient problem, which makes it difficult to learn long-range temporal dependencies. Long short-term memory (LSTM) networks and gated recurrent units (GRUs) were developed as practical RNN variants to address this limitation. Because of their ability to model temporal dependencies, RNN-based architectures have been widely applied to sequential neuroimaging data, such as fMRI time courses (Yan et al., 2019).

4. Encoder-Decoder

An encoder–decoder architecture consists of two main components: an encoder that maps input data into a latent representation and a decoder that transforms this representation into the desired output. This framework is widely used in tasks that require mapping one data structure to another, such as sequence-to-sequence applications in machine translation. Variational autoencoder (VAE) is a generative extension of the encoder–decoder framework. In a VAE, the encoder maps the input data to the parameters of a probability distribution, and the decoder samples from this latent distribution to generate new or reconstructed outputs (Aglinskas et al., 2022; Kim et al., 2021). Transformer modules can also be implemented within encoder–decoder architectures (Vaswani et al., 2017). However, many modern Transformer-based models use only one part of this framework depending on the task, including encoder-only models, such as BERT, and decoder-only models, such as ChatGPT.

5. Generative adversarial networks (GANs)

GANs are proposed as a framework for implicitly learning complex data distributions through adversarial learning. A GAN consists of two components: a generator, which produces synthetic samples, and a discriminator, which learns to distinguish generated samples from real data. During training, the discriminator provides feedback to the generator through gradient updates, encouraging the generator to produce increasingly realistic samples. Importantly, GANs are not a single model architecture but a general generative framework; different neural network architectures, such as CNNs or GCNs, can be used as the backbone of the generator and discriminator. In neuroimaging, representations learned by GANs have been applied to a range of tasks, including multi-site harmonization (Liu & Yap, 2024) and subtype identification (Yang et al., 2021). While GANs are highly effective at modeling complex data distributions, achieving optimal performance often benefits from sufficiently large datasets and thoughtfully designed architectures. Ongoing methodological advances continue to improve training stability and sample diversity, helping to address challenges such as mode collapse, in which the generator may produce samples that represent only a limited portion of the underlying data distribution or latent space (Gilpin, 2024).

6. Diffusion models

Similar to GANs, diffusion models represent a general generative modeling framework rather than a single fixed architecture. These models learn complex data distributions by gradually reversing a stochastic noise-adding process, with DL architectures such as U-Net commonly used to parameterize the denoising function. In neuroimaging, diffusion models have been applied to generate realistic fMRI signals, providing a useful approach for modeling temporal brain activity and brain dynamics (Hu et al., 2025). They have also been used to synthesize FC and augment fMRI datasets, thereby improving downstream analysis and disease classification performance (Zhao et al., 2025). Together, these studies suggest that diffusion models offer a promising generative framework for characterizing brain dynamics and investigating brain state transitions.

Given space limitations and the diverse backgrounds of readers, this chapter does not provide detailed explanations of DL concepts or their mathematical foundations. Readers seeking a comprehensive introduction to DL are referred to (Goodfellow et al., 2016). Sections 8.2 and 8.3 organize DL models into two broad categories: deep discriminative models, which learn conditional mappings from inputs to outputs, and deep generative models, which model the underlying data distribution for sample generation.

8.2. Deep discriminative models: learning conditional mappings from inputs to targets

Discriminative models learn the conditional mapping from inputs to outputs by modeling the conditional probability distribution $P(y|x)$, where x represent the input features and y denotes the target variables (*e.g.,* class labels, continuous outcomes). Standard discriminative machine learning models (*e.g.,* logistic regression, support vector machines) typically rely on manually engineered features or relatively simple transformations of the input space. In contrast, deep discriminative models use hierarchical feature-learning to capture complex nonlinear relationships in the data. By transforming high-dimensional inputs into low-dimensional, task-relevant latent representations, these models can create subspaces in which class boundaries or feature–outcome associations are easier to identify than in the original feature space. These learned representations can also support downstream analyses, including subtype identification through clustering in the latent space and biomarker discovery through post-hoc model interpretation.

8.2.1. Classification and regression

Classification and regression are among the most widely explored discriminative DL applications for brain dynamic analysis. The key distinction between them lies in the nature of the target variable: classification predicts discrete labels, whereas regression predicts continuous outcomes. In brain dynamic analysis, classification is commonly used for disease diagnosis or brain state identification, while regression is often applied to predict continues measures, such as disease risk, symptom severity, or cognitive scores.

Despite the ability of DL models to learn high-level feature representations from raw data, neuroimaging poses several intrinsic challenges. As shown in Table 1, neuroimaging data differ from natural images in important ways. Natural images are

typically two-dimensional and available in large datasets, whereas neuroimaging data are often high-dimensional, noisy, and limited in sample size. These characteristics create substantial obstacles for brain dynamics analysis. Therefore, although DL can automatically extract informative features, efficient feature processing or dimensionality reduction is still often recommended to reduce redundancy before modeling, particularly for MRI data with high dimensionality and relatively low signal-to-noise ratio (Yin et al., 2022).

For fMRI analysis, raw data are commonly reduced using atlas-based or data-driven approaches, such as independent component analysis (Calhoun et al., 2014). The resulting regional or component-level time courses can then be used to calculate static FC or dynamic FC. Because structural and functional MRI have distinct data characteristics, different DL architectures are typically selected for different neuroimaging modalities, as illustrated in **Figure 1**. Structural neuroimaging data, such as T1-weighted MRI and diffusion MRI, primarily capture voxel-wise tissue properties and structural connectivity. Since these modalities preserve rich three-dimensional spatial information, spatial feature extraction models, such as 3D CNNs, are commonly applied.

In contrast, fMRI has relatively lower spatial resolution but provides rich temporal information. Accordingly, fMRI can be modeled either as a sequence of time points or as a sequence of whole-brain volumes. When directly modeling fMRI signals, architectures that capture temporal dynamics, such as RNNs, or joint spatiotemporal patterns, such as CNN-RNN models, are commonly used. RNN-based models have been successfully applied to brain disorder diagnosis, brain decoding, and the detection of temporally dynamic functional state transitions (Rahman et al., 2022; Yan et al., 2018).

In scenarios with limited sample size, FC is often used as a more compact representation derived from fMRI. Unlike natural images, in which neighboring pixels or voxels have local spatial relationships, FC represents graph-like relationships among brain regions or networks. When FC matrices are flattened into one-dimensional feature vectors, can be used as a practical modeling approach. However, GNNs are theoretically better suited for FNC analysis because they can preserve the inherent topological structure of brain connectivity patterns (Bessadok et al., 2023).

A variety of deep discriminative models have been developed to capture spatiotemporal patterns in brain dynamics for classification and regression tasks. These models differ in how they represent temporal dynamics, spatial organization, and FC. For example, DSAM integrates temporal causal convolutional networks to model low- and high-level temporal features, a temporal attention unit to identify informative time points, a self-attention unit to construct task-specific connectivity matrices, and a graph neural network variant to capture spatial dynamics for downstream classification (Thapaliya et al., 2025). Similarly, the dFCExpert framework combines GNNs and a mixture-of-experts strategy to model dynamic FC for applications such as sex classification and autism identification (Chen et al., 2026). Multi-view architectures have also been proposed; for instance, MVF-XT uses cross-attention mechanisms to jointly model static and dynamic FC for obesity disorder classification (Fan et al., 2026), while hybrid models such as SD-CNN explicitly combine static and dynamic features for brain state classification(Huang et al., 2023). In regression settings, dynamic functional connectivity has been used for brain age prediction, as demonstrated by BrainATCL, which learns adaptive temporal connectivity patterns (Huang et al., 2025). Spatiotemporal attention

models have also shown promise for disease-risk prediction, such as detecting Alzheimer's-related alterations in asymptomatic individuals (Wei et al., 2024). Beyond connectivity-based representations, some models directly learn from fMRI time series; for example, convolutional layers can be used to capture inter-regional relationships, followed by recurrent neural networks to model temporal dependencies for schizophrenia classification (Yan et al., 2019).

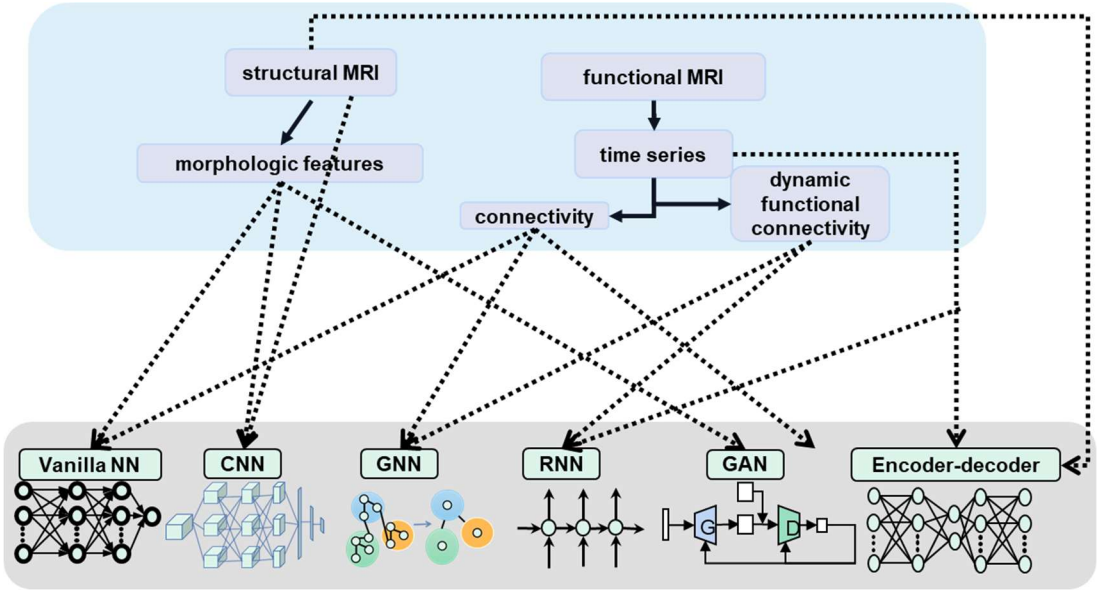

Figure 1. Different levels of MRI features and suitable DL approaches. In the gray panel, multiple deep learning modules are listed and linked with their applicable features. *Abbreviations*: CNN: convolutional neural network; GNN: graph neural network; RNN: recurrent neural network. GAN: generative adversarial network.

8.2.2. Subtype identification

Subtype identification is important for characterizing disease heterogeneity, advancing personalized medicine, and guiding targeted interventions. In neuroimaging studies, subtype discovery aims to identify patterns of brain structure or function that are consistently associated with variations in clinical symptoms, disease progression, or treatment responses. This goal is partly supported by the manifold hypothesis, which suggests that high-dimensional data may be organized along lower-dimensional manifolds that capture intrinsic structure. However, in neuroimaging analyses, clustering results can be influenced by confounding factors such as age, sex, or site effects. Therefore, effective subtype identification requires methods that can reduce the impact of confounds while uncovering biologically and clinically meaningful latent structure.

In SML-based subtype discovery, informative features often need to be carefully engineered before clustering. For example, in identifying subtypes of major depressive disorder based on FC, canonical component analysis (CCA) has been used to map FC features into a symptom-related subspace before applying clustering algorithms to identify distinct subtypes (Drysdale et al., 2017). This approach illustrates how feature transformation can help align neuroimaging variation with clinically relevant dimensions.

In contrast to SML models that rely on handcrafted features, DL models can automatically extract informative embeddings through multilayer architectures optimized by backpropagation. These learned representations may capture continuous biological variation even when the models are trained using discrete class labels. For example, as shown in **Figure 2a**, Abrol et al. (2021) trained a CNN to classify T1-weighted MRI data according to sex and age group using one-hot encoded labels. Visualization of the learned hidden-layer features with t-distributed stochastic neighbor embedding (t-SNE) revealed distinct clusters corresponding to sex, as well as a gradual progression along the age axis. (Abrol et al., 2021). Although the model was trained in a supervised classification setting, these findings suggest that DL can learn low-dimensional representations that reflect intrinsic biological structure beyond the original labels. This property is particularly relevant for subtype identification, as similar latent representations may be leveraged in unsupervised or weakly supervised settings to discover previously unknown subgroups within heterogeneous neuroimaging datasets. Comparable observations have also been reported in disease classification studies, such as the use of T1-weighted MRI to Huntington's disease classification (Plis et al., 2014).

The continuous spectrum learned by DL can be understood as low-dimensional manifolds that map high-dimensional input data into a task-relevant subspace. For example, when a DL model is trained to distinguish psychiatric disorders from healthy controls using fMRI time courses as inputs, it may transform the inputs into a psychiatric disorder–relevant representation while reducing the influence of less task-relevant factors, such as age or gender. As shown in **Figure 2b**, Yan et al. first trained a supervised multi-class classification model to project fMRI time series into a latent subspace in which differences among psychiatric disorders became more distinguishable. The high-level representations learned by the model were then visualized using t-SNE, revealing group-level separation across disorders and supporting the discovery of subtypes within schizoaffective disorder. This example illustrates how discriminative DL models can generate latent representations that not only improve classification but also reveal continuous or clustered structures relevant to disease heterogeneity. (Yan et al., 2022).

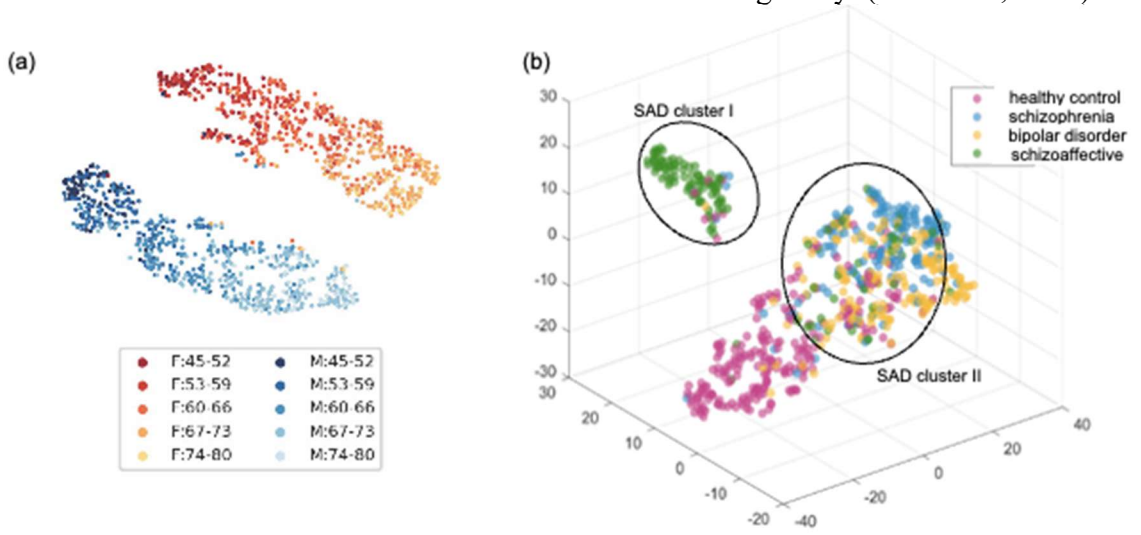

Figure 2. Subtype discovery using DL embeddings. **(a)** The MRI embeddings inferred from a DL model trained for age and gender classification. TSNE visualization of the DL hidden-layer features revealed clusters with a smooth progression along the age axis

(Abrol et al., 2021). **(b)** The fMRI embeddings inferred from a DL model trained for multi-class mental disorder classification. The DL not only revealed the relationships between mental disorders but also clustered schizoaffective disorder samples into two subtypes (black circled) (Yan et al., 2022).

8.2.3. Model interpretation for biomarker discovery

The goal of biomarker discovery is to identify interpretable features, such as the connectivity strength between two brain regions, that are reliably associated with target variables, such as the presence of schizophrenia. In DL-based neuroimaging studies, biomarker discovery often depends on model interpretation, which remains challenging because DL models use multiple nonlinear transformations to map input features into latent subspaces (Rahman et al., 2022). For a neuroimaging-based machine learning model to be interpretable, it should meet several requirements: it should be comprehensible to humans, provide meaningful information about the mental or behavioral constructs represented by specific brain regions or pathways, and demonstrate that its predictions are driven by relevant neurobiological signals rather than artifacts or confounds (Kohoutová et al., 2020). The growing need for interpretability has therefore motivated the development of various model introspection techniques. These techniques can be broadly divided into model-sensitive approaches, which depend on the internal structure of a specific model, and model-agnostic approaches, which can be applied independently of model architecture (Ribeiro et al., 2016). Each category has distinct advantages and limitations, and its suitability depends on the specific goals and requirements of the application (Wojciech Samek et al., 2021).

Model-sensitive interpretation relies on the internal structure or gradients of the trained model, and mainly includes gradient-based methods and layer-wise relevance propagation (LRP) (W. Samek et al., 2021). Gradient-based methods use automatic differentiation to estimate how changes in input features influence model outputs, and they can be applied without modifying the original DL architecture. For example, gradient-based interpretation has been used to identify brain regions associated with schizophrenia classification (Oh et al., 2019). However, these methods can be computationally expensive, particularly when higher integration precision requires a larger number of sampling steps. In contrast, LRP takes advantage of the layered structure of neural networks and propagates relevance scores backward through the model to generate explanations. Because LRP can be applied at the level of individual input samples, it supports interpretation across multiple levels of granularity, ranging from group-level patterns to single subjects, trials, or time points (Bach et al., 2015).

Model-agnostic interpretation does not depend on the internal architecture of the DL model. Instead, it typically evaluates model behavior by perturbing the input and measuring the resulting changes in the model output. Representative methods include occlusion sensitivity analysis, Local Interpretable Model-Agnostic Explanations (LIME), and Shapley Additive Explanations (SHAP). Occlusion analysis has been applied to CNN- and RNN-based models to estimate the contribution of individual brain regions to classification performance. For instance, Yan et al. (2019) trained a deep convolutional recurrent neural network to distinguish patients with schizophrenia from healthy controls using fMRI time courses extracted from 50 brain regions. To identify schizophrenia-related biomarkers, each brain region was iteratively occluded, and the resulting decrease

in classification performance was used to rank regional importance. This analysis identified the striatum as one of the most informative regions for schizophrenia classification. (Yan et al., 2019). LIME explains a DL model by approximating its behavior locally around a given input sample using a simpler interpretable model, such as a linear model. SHAP estimates feature importance using Shapley values, which quantify the contribution of each feature across possible feature subsets. For example, Lombardi et al. (2021) applied both SHAP and LIME to assess the contribution of brain morphological descriptors to predicted brain age and reported that SHAP provided more reliable explanations of morphological aging patterns (Lombardi et al., 2021).

8.3. Deep generative models: learning data distributions for generating new samples

Deep generative models aim to model the underlying data distribution by learning either the marginal distribution *P(x)* or the joint probability distribution *P(x,y)*, thereby enabling the synthesis of new samples that approximate the observed data. Unlike conventional generative models, which often depend on explicit probabilistic assumptions, deep generative models use neural networks to capture complex and high-dimensional data distributions (Ramezanian-Panahi et al., 2022). Representative deep generative models include VAEs, GANs, and diffusion models. VAEs learn latent representations through probabilistic encoder–decoder architectures that approximate the data distribution (Aglinskas et al., 2022; Kim et al., 2021). GANs use an adversarial training framework in which a generator and a discriminator are jointly optimized to produce realistic samples (Creswell et al., 2018; Gilpin, 2024). Diffusion models generate data by learning to reverse a stochastic noise process, providing stable training and strong coverage of complex data distributions (Hu et al., 2025; Zhao et al., 2025). In neuroimaging, these models are particularly valuable because they can capture nonlinear interactions, model latent structure, and disentangle confounding factors such as demographic variability and scanner effects. These properties make deep generative models useful for characterizing brain dynamics, including temporal variability and evolving FC patterns, while also supporting applications such as subtype discovery, data augmentation, multi-site harmonization, and latent brain state analysis.

8.3.1. Subtype identification

Brain development and brain disorders are shaped by dynamic interactions between genetic, environmental, and disease-related factors. In neuroimaging studies, subtype discovery is further complicated by confounding variables such as scanner effects, age, and gender, which may obscure disease-relevant heterogeneity if not properly accounted for. As discussed in Section 8.2.2, deep discriminative models can map input features into task-related subspaces and reveal low-dimensional manifolds that are useful for subtype discovery. However, because these models are primarily optimized for prediction or classification, they often have limited ability to explicitly disentangle disease-relevant variation from confounding or background variability.

Deep generative models provide an alternative framework by explicitly modeling data distributions and separating different sources of variation in the latent space. Here, we introduce two representative generative approaches for MRI-based subtype discovery: contrastive variational autoencoders (CVAEs) and Smile-GAN. CVAEs have been

increasingly used in neuroimaging to disentangle shared and condition-specific variation. For example, Aglinskas et al. (2022) applied CVAE to identify autism spectrum disorder subtypes by isolating disorder-specific neuroanatomical features from features shared with healthy controls (Aglinskas et al., 2022). This framework has also been extended to brain dynamics. Ding et al. (2024) used CVAE to extract disorder-specific neurodynamic representations in schizophrenia (Ding et al., 2024). In this model, data from both healthy controls and patients are decomposed into two latent spaces: one capturing shared variation across groups and the other capturing disease-specific variation. By separating schizophrenia-related dynamic patterns from background variability, the model generates latent features that are more directly related to disease heterogeneity. These disease-specific representations were significantly associated with clinical measures, suggesting that they capture biologically meaningful variation. Although clustering was not explicitly performed in this study, the learned latent space provides a natural basis for subtype discovery, as downstream clustering or mixture modeling can be applied to identify subgroups based on disorder-specific features.

CVAE-based subtype discovery therefore typically follows a two-stage procedure: first, disentangling disease-relevant features from shared or confounding variation, and second, applying clustering methods to the learned latent representations. A more direct strategy is to integrate clustering into the generative modeling framework itself. Because few studies have combined GAN-based models with dynamic functional features for subtype discovery, we use Smile-GAN as a representative structural MRI-based example to illustrate this idea. As shown in Figure 3b, Smile-GAN learns nonlinear mappings from healthy controls to patients, allowing the model to separate disease-relevant from disease-irrelevant variation. In addition, it incorporates a clustering module into the training process, enabling subjects to be grouped based on the learned disease-related features. By jointly optimizing the generator, discriminator, and clustering components, Smile-GAN not only models realistic patient-specific variations but also directly identifies biologically meaningful subgroups, providing a unified framework for generative modeling and subtype discovery (Yang et al., 2021).

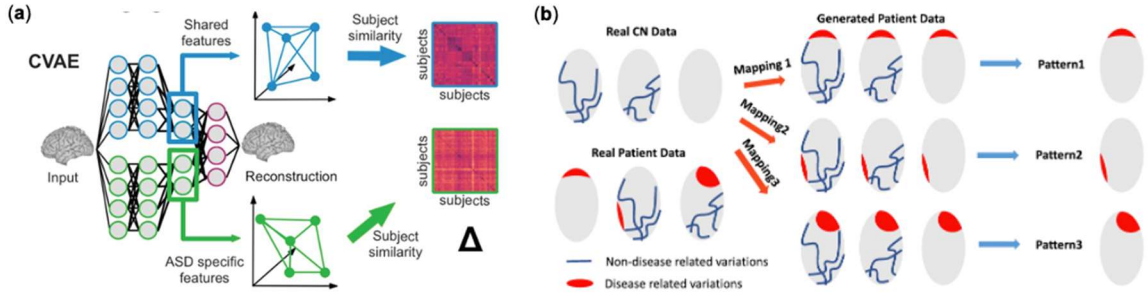

Figure 3. (**a**) Architecture of the CVAE model. CVAEs take input samples from two distinct populations and isolate variation specific to one population from common variation. As a result, CVAEs disentangle "autism-specific" neuroanatomical variation from variation "shared" by both autism and healthy controls, representing each as a distinct set of latent features (**b**) Conceptual overview of Smile-GAN. Blue curves represent non-disease-related variations observed in both healthy controls and patients. Red regions represent disease-specific neuroanatomical patterns by learning non-linear transformations from normal controls to subtypes.

8.3.2. Data augmentation and multi-site dataset harmonization

Brain dynamics analysis is often constrained by limited data availability, which motivates the use of data augmentation and large-scale data collaboration. Although pooling data across multiple sites can help address small sample issues, it introduces additional challenges because differences in scanners, acquisition protocols, and preprocessing pipelines can create site-specific biases that obscure true biological signals and hinder model generalization. Deep generative models offer a promising solution by learning the underlying distribution of neuroimaging data, generating realistic synthetic samples, and separating site-related variation from biologically meaningful signals.

One important application of deep generative models is data augmentation for fMRI-based brain dynamics analysis. Recent studies have explored GANs, VAEs, and related generative frameworks to synthesize fMRI signals or FC, thereby alleviating data scarcity and improving model generalization (Zhuang et al., 2019). For example, VAE-GAN frameworks have been proposed to jointly model temporal features and functional brain networks while generating high-quality synthetic fMRI data. These synthetic data can enhance downstream tasks such as classification and network identification (Qiang et al., 2023). Together, these studies suggest that generative models can capture meaningful spatiotemporal structure in brain data and provide effective augmentation for dynamic analysis.

Another important application is multi-site harmonization. A successful harmonization model should either estimate the distribution of site-specific variation or disentangle site-related effects from biologically relevant image features. GAN-based methods have been widely used for this purpose (Bashyam et al.; Liu & Yap, 2024; Yan et al., 2023). For instance, Bashyam et al. applied a GAN variant, StarGAN, to harmonize T1-weighted MRI data collected from six sites (Bashyam et al., 2022). The model consists of a style encoder, a content encoder, a generator, and a discriminator. The style encoder maps each axial slice to an eight-dimensional representation of site-related variation, whereas the content encoder extracts site-invariant anatomical information using convolutional filters. The generator then combines the style and content representations to produce a harmonized image that matches the site-related characteristics of a reference scan while preserving the anatomical content of the original image. The discriminator further encourages the generated images to appear realistic. This framework illustrates how generative models can support multi-site collaboration by reducing scanner- or site-related confounds while maintaining biologically meaningful information.

8.3.3. Generative latent state analysis of brain dynamics

Brain dynamic analysis is typically studied through two related but distinct representations: dynamic FC and dynamic brain activity. Dynamic FC characterizes time-varying changes in brain network organization, whereas dynamic brain activity directly models temporal fluctuations in neural signals, such as fMRI BOLD activity, EEG recordings, or neural spiking activity. Deep generative models provide a useful framework for both types of analysis because they can reconstruct observed brain signals while learning the intrinsic data distribution. In doing so, they map high-dimensional neural data into lower-dimensional latent representations that can be used to characterize

brain state transitions, temporal trajectories, and the evolution of network structure over time.

In the context of dynamic FC, Gomez et al. (2024) proposed a low-dimensional VAE framework to model dFC associated with different states of consciousness. The model projected dynamic FC into a two-dimensional latent space, where brain connectivity patterns and consciousness conditions were organized into interpretable trajectories. This study further showed that generative latent representations can be used to simulate transitions between wakefulness and unconsciousness (Gomez et al., 2024). Similarly, Campbell et al. (2024) developed the Dynamic Brain Graph Deep Generative Model (DBGDGM), which represents fMRI-derived dynamic brain graphs as fixed brain regions connected by time-varying edges defined by dFC. DBGDGM learns graph-, node-, and community-level embeddings in an unsupervised manner, allowing brain regions to be grouped into temporally evolving communities. By sampling node embeddings from time-varying community distributions, the model captures dynamic reorganization of functional brain networks rather than relying on static connectivity patterns. Its superior performance in graph reconstruction and dynamic link prediction, together with the meaningful overlap between learned communities and known functional connectivity patterns, suggests that deep generative graph models can reveal interpretable patterns of evolving brain connectivity (Campbell et al., 2024).

Dynamic brain activity analysis, in contrast, focuses on recovering time-varying neural population dynamics from high-dimensional signals such as neural spiking activity, fMRI BOLD signals, or EEG recordings. Traditional approaches, such as linear dynamical systems (LDS), often assume that latent neural dynamics are linear, independent, or piecewise linear, which limits their ability to capture transient nonlinear patterns related to cognition, behavior, or disease. Deep generative models address this limitation by learning low-dimensional latent dynamical systems that explain the observed neural signals. For example, as shown in Figure 4a, latent factor analysis via dynamical systems (LFADS) uses a sequential variational autoencoder to infer single-trial latent trajectories from neural spiking data, supporting denoising, firing-rate estimation, behavioral prediction, and characterization of trial-to-trial variability (Pandarinath et al., 2018). Similarly, as shown in Figures 4b–c, generative state-space models based on piecewise-linear recurrent neural networks (PLRNNs) have been applied to fMRI data to uncover task-related nonlinear dynamics that are not captured by conventional linear models (Koppe et al., 2019).

Together, these studies demonstrate that deep generative models can support brain dynamics analysis at both the connectivity and activity levels. By reconstructing observed data while learning compact latent representations, these models can reveal latent neural dynamics that link brain signals to behavior, task processing, and clinical states.

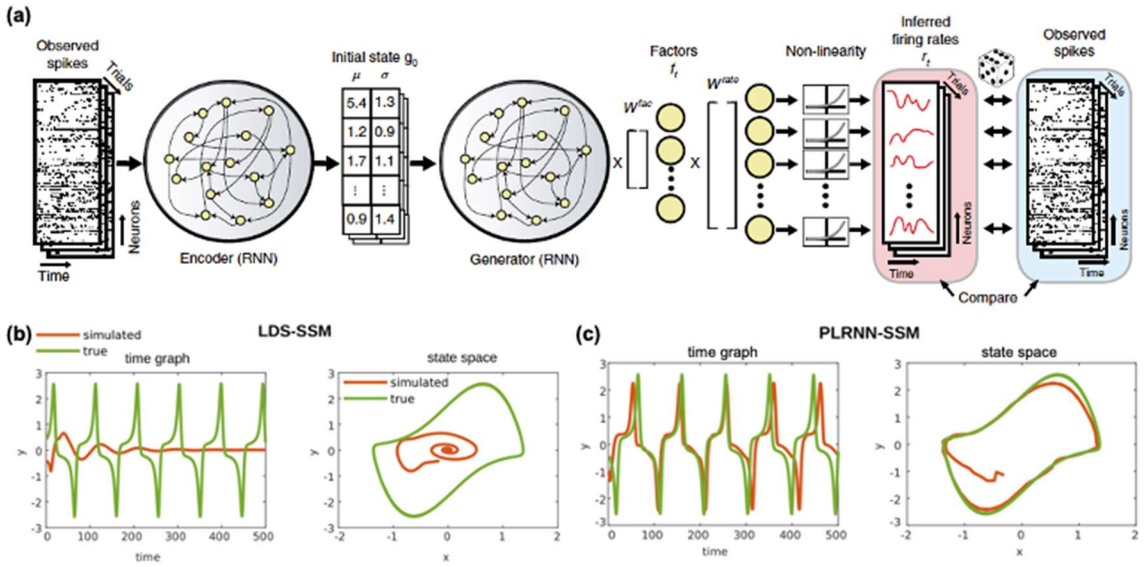

Figure 4. (**a**) schematic overview of the LFADS architecture (Pandarinath et al., 2018). (**b**) Example time series from an LDS-SSM and a PLRNN-SSM trained on the vdP system. Example time graph (left) and state space (right) for a trajectory generated by an LDS-SSM (red) trained on the vdP system (true vdP trajectories 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 (Koppe et al., 2019).

8.4. Future studies

8.4.1. The balance between task complexity and model complexity

As DL models are increasingly applied to brain dynamic analysis, an important challenge is the trade-off between task complexity, model complexity, and available training samples. In general, DL models applied to more abstract features representations (*e.g.,* FC) require less complexity than those trained on raw features (*e.g.,* fMRI time series). Likely, simpler tasks (*e.g.,* sex classification) can often be addressed with DL models with fewer trainable parameters (*e.g.,* CNN, GCNs) than more complex tasks (*e.g.,* brain state identification). Although auto-differentiation platforms (*e.g.,* Pytorch, TensorFlow) have greatly simplified model design procedures, various hyper-parameters such as width, depth, loss function, and optimizers remain heavily dependent to experiences, creating a gap between practical model development and theoretical understanding. In brain dynamic analysis, different tasks require different performance criteria. For example, a DL model used for psychiatric risk prediction may require higher sensitivity and robustness than a model used for an exploration of mental states. Future studies should develop theoretical and empirical frameworks for estimating the minimal model complexity, feature representation, and sample size, for a given task. Such frameworks would help guild the selection of efficient architectures for brain dynamics and provide quantitative basis for understanding the DL outputs.

8.4.2. Neuromodulation and multi-level brain dynamic analysis

Neuromodulatory systems play a key role in regulating brain state transitions, dynamic FC, and cognitive flexibility, making them critical for understanding how brain dynamics emerge from molecular and cellular mechanisms (Kringelbach et al., 2020). A comprehensive understaning of brain dynamics therefore requires multilayered computaitonal models that bridge molecular, neural, systems, and behavioral levels of analysis (Shine et al., 2019; Shine et al., 2021). In this context, integrating molecular-level measures, such as PET-derived neurotransmitter or receptor maps, with system-level measures, such as fMRI activity and dynamic FC, may provide a more mechanistic interpretation of brain dynamics than approaches based solely on structural or functional connectivity. Incorporating DL into neuromodulatory modeling could further enhance these models by capturing nonlinear relationships among molecular neuromodulation, dynamic brain states, and behavioral outcomes. This interdisciplinary framework may ultimately support the development of models that not only mimic brain-like computation but also reveal how neuromodulatory processes shape temporal variability, state transitions, and cognitive function.

8.4.3. Foundation models for brain dynamics

Foundation models are large-scale, pretrained architectures that extract universal features from diverse brain imaging modalities using self-supervised and contrastive learning features. They utilize advanced methods like masked autoencoding, graph representations, and transformers to enable rapid adaptation to tasks. By leveraging large-scale pretraining and transferable representations, foundation models can improve sample efficiency and enable downstream models to be fine-tuned with fewer labeled subjects. Recent neuroimaging foundation models have shown strong potential in segmentation (Cox et al., 2024), MRI enhancement (Sun et al., 2025), registration (Wang et al., 2025), diagnosis (Sun et al., 2025), and biomarker discovery (Tak et al., 2026). More directly related to brain dynamics, Brain-JEPA uses spatiotemporal masking and brain gradient positioning to learn transferable representations from fMRI data, supporting tasks such as demographic prediction, disease diagnosis, prognosis, and trait prediction (Dong et al., 2024). However, most of the existing foundation models are pretrained from imaging features and rarely incorporate biological information (*e.g.,* neurotransmitter systems, gene expression) during pretraining, potentially limiting interpretability, mechanistic insight, and generalizability across populations and disease conditions. Future work should therefore move beyond purely image-based representation learning toward biologically informed foundation models that jointly capture spatial organization, temporal evolution, and multimodal biological mechanisms.

Acknowledgements

This research is supported by the Intramural Research Program of the National Institutes of Health. The author thanks Dr. Nora D. Volkow for her valuable comments and suggestions. Portions of the text were polished with the assistance of artificial intelligence tools, with all content reviewed and edited by the author.

References

Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., & Calhoun, V. (2021). Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature Communications*, *12*(1), 353. https://doi.org/10.1038/s41467-020-20655-6

Aglinskas, A., Hartshorne, J. K., & Anzellotti, S. (2022). Contrastive machine learning reveals the structure of neuroanatomical variation within autism. *Science*, *376*(6597), 1070-1074. https://doi.org/doi:10.1126/science.abm2461

Bach, S., Binder, A., Montavon, G., Klauschen, F., Muller, K. R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS One*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Bashyam, V. M., Doshi, J., Erus, G., Srinivasan, D., Abdulkadir, A., Singh, A., Habes, M., Fan, Y., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Satterthwaite, T. D., Wolf, D. H., Gur, R. E., Gur, R. C., . . . consortia, P. Deep Generative Medical Image Harmonization for Improving Cross-Site Generalization in Deep Learning Predictors. *Journal of Magnetic Resonance Imaging*, *n/a*(n/a). https://doi.org/https://doi.org/10.1002/jmri.27908

Bashyam, V. M., Doshi, J., Erus, G., Srinivasan, D., Abdulkadir, A., Singh, A., Habes, M., Fan, Y., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Satterthwaite, T. D., Wolf, D. H., Gur, R. E., Gur, R. C., . . . consortia, P. (2022). Deep Generative Medical Image Harmonization for Improving Cross-Site Generalization in Deep Learning Predictors. *Journal of Magnetic Resonance Imaging*, *55*(3), 908-916. https://doi.org/https://doi.org/10.1002/jmri.27908

Bessadok, A., Mahjoub, M. A., & Rekik, I. (2023). Graph Neural Networks in Network Neuroscience. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(5), 5833-5848. https://doi.org/10.1109/TPAMI.2022.3209686

Calhoun, V. D., Miller, R., Pearlson, G., & Adali, T. (2014). The Chronnectome: Time-Varying Connectivity Networks as the Next Frontier in fMRI Data Discovery. *Neuron*, *84*(2), 262-274. https://doi.org/10.1016/j.neuron.2014.10.015

Campbell, A., Spasov, S. E., Toschi, N., & Lio, P. (2024). *DBGDGM: Dynamic Brain Graph Deep Generative Model* Medical Imaging with Deep Learning, Proceedings of Machine Learning Research. https://proceedings.mlr.press/v227/campbell24b.html

Chen, T., Li, H., Zheng, H., & Fan, Y. (2026). dFCExpert: Learning Dynamic Functional Connectivity Patterns With Modularity and State Experts. *IEEE Transactions on Medical Imaging*, *45*(3), 1088-1098. https://doi.org/10.1109/TMI.2025.3617310

Cox, J., Liu, P., Stolte, S. E., Yang, Y., Liu, K., See, K. B., Ju, H., & Fang, R. (2024). BrainSegFounder: Towards 3D foundation models for neuroimage segmentation. *Medical Image Analysis*, *97*, 103301. https://doi.org/https://doi.org/10.1016/j.media.2024.103301

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*, *35*(1), 53-65. https://doi.org/10.1109/msp.2017.2765202

Ding, C., Sun, Y., Li, K., Xie, S., Yan, H., Li, P., Yan, J., Chen, J., Wang, H., Wang, H., Chen, Y., Yang, Y., Lv, L., Zhang, H., Lu, L., Zhang, D., Chen, Y., Zhang, Z., Jiang, T., & Liu, B. (2024). Disorder-specific neurodynamic features in schizophrenia inferred by neurodynamic embedded contrastive variational autoencoder model. *Translational Psychiatry*, *14*(1), 496. https://doi.org/10.1038/s41398-024-03200-7

Dong, Z., Li, R., Wu, Y., Nguyen, T. T., Chong, J. S., Ji, F., Tong, N. R., Chen, C. L., & Zhou, J. H. (2024). Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in neural information processing systems*, *37*, 86048-86073.

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R. N., Zebley, B., Oathes, D. J., Etkin, A., Schatzberg, A. F., Sudheimer, K., Keller, J., Mayberg, H. S., Gunning, F. M., Alexopoulos, G. S., Fox, M. D., Pascual-Leone, A., Voss, H. U., . . . Liston, C. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat Med*, *23*(1), 28-38. https://doi.org/10.1038/nm.4246

Fan, X., Li, J., Huang, H., Li, G., Zhang, W., Hu, Y., Sun, W., Yan, W., Manza, P., Volkow, N. D., Wang, G.-J., & Zhang, Y. (2026). MVF-XT: An interpretable multi-view fusion network based on cross-attention for fMRI analysis. *Neurocomputing*, *677*, 133108. https://doi.org/https://doi.org/10.1016/j.neucom.2026.133108

Gilpin, W. (2024). Generative learning for nonlinear dynamics. *Nature Reviews Physics*. https://doi.org/10.1038/s42254-024-00688-2

Gomez, C., Uhrig, L., Frouin, V., Duchesnay, E., Jarraya, B., & Grigis, A. (2024). Deep learning models reveal the link between dynamic brain connectivity patterns and states of consciousness. *Scientific Reports*, *14*(1), 31606. https://doi.org/10.1038/s41598-024-76695-1

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Han, K., Su, Y., He, L., Zhan, L., Plis, S., Calhoun, V., & Yang, C. (2026). Rethinking functional brain connectome analysis: do graph deep learning models Help. *npj Artificial Intelligence*, *2*(1). https://doi.org/10.1038/s44387-025-00067-x

He, K., Zhang, X., Ren, S., & Sun, J. (2016, 27-30 June 2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),

Hu, Y., Li, W., & Yuan, Y. (2025). Synthesizing realistic fMRI: a physiological dynamics-driven hierarchical diffusion model for efficient fmri acquisition. The Thirteenth International Conference on Learning Representations,

Huang, J., Wang, M., Ju, H., Shi, Z., Ding, W., & Zhang, D. (2023). SD-CNN: A static-dynamic convolutional neural network for functional brain networks. *Medical Image Analysis*, *83*, 102679. https://doi.org/https://doi.org/10.1016/j.media.2022.102679

Huang, Y., Nouranizadeh, A., Ahrends, C., & Xu, M. (2025). BrainATCL: Adaptive Temporal Brain Connectivity Learning for Functional Link Prediction and Age Estimation. *arXiv preprint arXiv:2508.07106*.

Kim, J.-H., Zhang, Y., Han, K., Wen, Z., Choi, M., & Liu, Z. (2021). Representation learning of resting state fMRI with variational autoencoder. *Neuroimage*, *241*, 118423. https://doi.org/https://doi.org/10.1016/j.neuroimage.2021.118423

Kim, J., Calhoun, V. D., Shim, E., & Lee, J. H. (2015). 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*(Pt A), 127-146. https://doi.org/10.1016/j.neuroimage.2015.05.018

Kohoutová, L., Heo, J., Cha, S., Lee, S., Moon, T., Wager, T. D., & Woo, C.-W. (2020). Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nature Protocols*, *15*(4), 1399-1435. https://doi.org/10.1038/s41596-019-0289-5

Koppe, G., Toutounji, H., Kirsch, P., Lis, S., & Durstewitz, D. (2019). Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fMRI. *PLOS Computational Biology*, *15*(8), e1007263. https://doi.org/10.1371/journal.pcbi.1007263

Kringelbach, M. L., Cruzat, J., Cabral, J., Knudsen, G. M., Carhart-Harris, R., Whybrow, P. C., Logothetis, N. K., & Deco, G. (2020). Dynamic coupling of whole-brain neuronal and neurotransmitter systems. *Proc Natl Acad Sci U S A*, *117*(17), 9566-9576. https://doi.org/10.1073/pnas.1921475117

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

Liu, S., & Yap, P.-T. (2024). Learning multi-site harmonization of magnetic resonance images without traveling human phantoms. *Communications Engineering*, *3*(1), 6. https://doi.org/10.1038/s44172-023-00140-w

Lombardi, A., Diacono, D., Amoroso, N., Monaco, A., Tavares, J. M. R. S., Bellotti, R., & Tangaro, S. (2021). Explainable Deep Learning for Personalized Age Prediction With Brain Morphology [Original Research]. *Frontiers in neuroscience*, *15*(578). https://doi.org/10.3389/fnins.2021.674055

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: a survey. *Multimedia Tools and Applications*, *80*(6), 9411-9457. https://doi.org/10.1007/s11042-020-10073-7

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. *Schizophrenia Research*, *212*, 186-195. https://doi.org/https://doi.org/10.1016/j.schres.2019.07.034

Pandarinath, C., O'Shea, D. J., Collins, J., Jozefowicz, R., Stavisky, S. D., Kao, J. C., Trautmann, E. M., Kaufman, M. T., Ryu, S. I., Hochberg, L. R., Henderson, J. M., Shenoy, K. V., Abbott, L. F., & Sussillo, D. (2018). Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, *15*(10), 805-815. https://doi.org/10.1038/s41592-018-0109-9

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., Johnson, H. J., Paulsen, J. S., Turner, J. A., & Calhoun, V. D. (2014). Deep

learning for neuroimaging: a validation study. *Front Neurosci*, *8*, 229. https://doi.org/10.3389/fnins.2014.00229

Qiang, N., Gao, J., Dong, Q., Yue, H., Liang, H., Liu, L., Yu, J., Hu, J., Zhang, S., Ge, B., Sun, Y., Liu, Z., Liu, T., Li, J., Song, H., & Zhao, S. (2023). Functional brain network identification and fMRI augmentation using a VAE-GAN framework. *Computers in Biology and Medicine*, *165*, 107395. https://doi.org/https://doi.org/10.1016/j.compbiomed.2023.107395

Rahman, M. M., Calhoun, V., & Plis, S. (2026). Deep learning interpretability in neuroimaging: A comprehensive survey and methodological recommendations. *Imaging Neurosci (Camb)*, *4*. https://doi.org/10.1162/IMAG.a.1129

Rahman, M. M., Mahmood, U., Lewis, N., Gazula, H., Fedorov, A., Fu, Z., Calhoun, V. D., & Plis, S. M. (2022). Interpreting models interpreting brain dynamics. *Scientific Reports*, *12*(1), 12023. https://doi.org/10.1038/s41598-022-15539-2

Ramezanian-Panahi, M., Abrevaya, G., Gagnon-Audet, J.-C., Voleti, V., Rish, I., & Dumas, G. (2022). Generative Models of Brain Dynamics [Review]. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.807406

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, *109*(3), 247-278.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proceedings of the IEEE*, *109*(3), 247-278. https://doi.org/10.1109/JPROC.2021.3060483

Shine, J. M., Breakspear, M., Bell, P. T., Ehgoetz Martens, K. A., Shine, R., Koyejo, O., Sporns, O., & Poldrack, R. A. (2019). Human cognition involves the dynamic integration of neural activity and neuromodulatory systems. *Nature Neuroscience*, *22*(2), 289-296. https://doi.org/10.1038/s41593-018-0312-0

Shine, J. M., Müller, E. J., Munn, B., Cabral, J., Moran, R. J., & Breakspear, M. (2021). Computational models link cellular mechanisms of neuromodulation to large-scale neural dynamics. *Nature Neuroscience*, *24*(6), 765-776. https://doi.org/10.1038/s41593-021-00824-6

Sun, Y., Wang, L., Li, G., Lin, W., & Wang, L. (2025). A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. *Nature Biomedical Engineering*, *9*(4), 521-538. https://doi.org/10.1038/s41551-024-01283-7

Tak, D., Garomsa, B. A., Zapaishchykova, A., Chaunzwa, T. L., Climent Pardo, J. C., Ye, Z., Zielke, J., Ravipati, Y., Pai, S., Vajapeyam, S., Mahootiha, M., Parker, M., Pike, L. R. G., Smith, C., Familiar, A. M., Liu, K. X., Prabhu, S., Arnaout, O., Bandopadhayay, P., . . . Kann, B. H. (2026). A generalizable foundation model for analysis of human brain MRI. *Nature Neuroscience*. https://doi.org/10.1038/s41593-026-02202-6

Tang, J., Zhu, T., Zhou, W., & Zhao, W. (2026). Graph neural networks for fMRI functional brain networks: A survey. *Neural Networks*, *194*, 108137. https://doi.org/https://doi.org/10.1016/j.neunet.2025.108137

Thapaliya, B., Miller, R., Chen, J., Wang, Y. P., Akbas, E., Sapkota, R., Ray, B., Suresh, P., Ghimire, S., Calhoun, V. D., & Liu, J. (2025). DSAM: A deep learning framework for analyzing temporal and spatial dynamics in brain networks. *Medical Image Analysis*, *101*, 103462. https://doi.org/https://doi.org/10.1016/j.media.2025.103462

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, E. Y., Fahey, P. G., Ding, Z., Papadopoulos, S., Ponder, K., Weis, M. A., Chang, A., Muhammad, T., Patel, S., Ding, Z., Tran, D., Fu, J., Schneider-Mizell, C. M., da Costa, N. M., Reid, R. C., Collman, F., da Costa, N. M., Franke, K., Ecker, A. S., . . . Consortium, M. I. (2025). Foundation model of neural activity predicts response to new stimulus types. *Nature*, *640*(8058), 470-477. https://doi.org/10.1038/s41586-025-08829-y

Wei, Y., Abrol, A., Lah, J., Qiu, D., & Calhoun, V. D. (2024, 15-19 July 2024). A deep spatio-temporal attention model of dynamic functional network connectivity shows sensitivity to Alzheimer's in asymptomatic individuals. 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122-1136. https://doi.org/10.1109/jas.2023.123618

Xu, M., Calhoun, V., Jiang, R., Yan, W., & Sui, J. (2021). Brain imaging-based machine learning in autism spectrum disorder: methods and applications. *Journal of Neuroscience Methods*, *361*, 109271. https://doi.org/https://doi.org/10.1016/j.jneumeth.2021.109271

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., . . . Sui, J. (2019). Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site FMRI data. *EBioMedicine*, *47*, 543-552. https://doi.org/10.1016/j.ebiom.2019.08.023

Yan, W., Fu, Z., Jiang, R., Sui, J., & Calhoun, V. D. (2023). Maximum Classifier Discrepancy Generative Adversarial Network for Jointly Harmonizing Scanner Effects and Improving Reproducibility of Downstream Tasks. *IEEE Transactions on Biomedical Engineering*, 1-9. https://doi.org/10.1109/TBME.2023.3330087

Yan, W., Qu, G., Hu, W., Abrol, A., Cai, B., Qiao, C., Plis, S. M., Wang, Y. P., Sui, J., & Calhoun, V. D. (2022). Deep Learning in Neuroimaging: Promises and challenges. *IEEE Signal Processing Magazine*, *39*(2), 87-98. https://doi.org/10.1109/MSP.2021.3128348

Yan, W., Zhang, H., Sui, J., & Shen, D. (2018). Deep Chronnectome Learning via Full Bidirectional Long Short-Term Memory Networks for MCI Diagnosis. *Med*

Image Comput Comput Assist Interv, *11072*, 249-257. https://doi.org/10.1007/978-3-030-00931-1_29

Yang, Z., Nasrallah, I. M., Shou, H., Wen, J., Doshi, J., Habes, M., Erus, G., Abdulkadir, A., Resnick, S. M., Albert, M. S., Maruff, P., Fripp, J., Morris, J. C., Wolk, D. A., Davatzikos, C., i, S. C., Baltimore Longitudinal Study of, A., & Alzheimer's Disease Neuroimaging, I. (2021). A deep learning framework identifies dimensional representations of Alzheimer's Disease from brain structure. *Nat Commun*, *12*(1), 7065. https://doi.org/10.1038/s41467-021-26703-z

Yin, W., Li, L., & Wu, F.-X. (2022). Deep learning for brain disorder diagnosis based on fMRI images. *Neurocomputing*, *469*, 332-345. https://doi.org/https://doi.org/10.1016/j.neucom.2020.05.113

Zhao, H., Lou, H., Yao, L., & Zhang, Y. (2025). Diffusion transformer-augmented fMRI functional connectivity for enhanced autism spectrum disorder diagnosis. *J Neural Eng*, *22*(1). https://doi.org/10.1088/1741-2552/adb07a

Zhuang, P., Schwing, A. G., & Koyejo, O. (2019, 8-11 April 2019). FMRI Data Augmentation Via Synthesis. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019),