Weizheng Yan, Gang Qu, Wenxing Hu, Anees Abrol, Biao Cai, Chen Qiao, Sergey M. Plis, Yu-Ping Wang, Jing Sui, and Vince D. Calhoun

Deep Learning in Neuroimaging

Promises and challenges

SHUTTERSTOCK.COM/KKSSR

Digital Object Identifier 10.1109/MSP.2021.3128348
Date of current version: 24 February 2022

Deep learning (DL) has been extremely successful when applied to the analysis of natural images. By contrast, analyzing neuroimaging data presents some unique challenges, including higher dimensionality, smaller sample sizes, multiple heterogeneous modalities, and a limited ground truth. In this article, we discuss DL methods in the context of four diverse and important categories in the neuroimaging field: classification/prediction, dynamic activity/connectivity, multimodal fusion, and interpretation/visualization. We highlight recent progress in each of these categories, discuss the benefits of combining data characteristics and model architectures, and derive guidelines for the use of DL in neuroimaging data. For each category, we also assess promising applications and major challenges to overcome. Finally, we discuss future directions of neuroimaging DL for clinical applications, a topic of great interest, touching on all four categories.

Introduction

Neuroimaging is a powerful tool that is being used to provide important insights into both healthy and disordered human brains. It also has the potential to translate discoveries and technological advances into the effective diagnosis, prevention, and treatment of brain disorders (https://braininitiative.nih.gov/). Flourishing neuroimaging techniques, such as magnetic resonance imaging (MRI) and magnetoencephalography (MEG), have revolutionized our ability to noninvasively study the human brain structure, function, wiring, and metabolism. In contrast to natural images, which are collected under natural light, neuroimaging data consist mostly of radiological images. Because of this, the noise distribution of neuroimaging varies depending on the acquisition used [e.g., Rician noise in MRI, quantum noise in computed tomography (CT)]. As shown in Table 1, neuroimaging data come with many other additional unique aspects, including the number of modalities, high dimensionality, low signal-to-noise ratio, and small sample sizes compared to natural image data.

Studies in neuroimaging using DL models initially appeared in 2014 [1], and the number of studies has rapidly grown since then, fueled by many new models as well as the accumulation of available data actively supported by various consortia and

funders (e.g., the Human Connectome Project, Alzheimer's Disease Neuroimaging Initiative, Enhancing NeuroImaging Genetics Through Meta-Analysis, Autism Brain Imaging Data Exchange, Adolescent Brain Cognitive Development, and UK Biobank). MRI, as a noninvasive technique with high spatiotemporal resolution, is currently the most widely studied neuroimaging modality according to the search terms used (Figure 1).

Advanced neuroimaging analysis approaches are essential for linking brain function and structure to network and behavior. Linear models and, in particular, flexible matrix decomposition approaches have contributed a lot to our current understanding. For instance, group independent component analysis (ICA), as a purely data-driven algorithm that reveals large-scale networks by making group inferences from functional MRI (fMRI), is particularly useful for data fusion of multiple modalities, such as genome-wide single-nucleotide polymorphism (SNP) data or event-related potentials [2]. Despite this, classical neuroimaging analytic approaches with standard machine learning (SML) methods have relatively limited model flexibility. SMLs often require considerable domain expertise to design feature extractors that can transform raw data into suitable internal representations or feature vectors from which the learning subsystems can detect or classify patterns [3]. Such "shallow" combinations of raw features can be sensitive to irrelevant variations and may not be flexible enough for revealing high-level differences or predicting complex brain–behavior relationships.

By contrast, DL uses multiple processing layers to learn representations of data with multiple levels of abstraction. Compared to SML, DL approaches are highly flexible and use minimally preengineered features. Though complex models are susceptible to "black-box" problems, representative features can now be learned automatically via different procedures to improve interpretability. Consequently, DL has turned out to be efficient in discovering intrinsic structure from high-dimensional data. Historically, breakthroughs often happen when data are relatively abundant, such as in text and natural image classification. As high-quality neuroimaging data sets accumulate, the performance of DL in neuroimaging will undoubtedly be significantly improved, and the combination of unsupervised models has the potential of making important advances in our understanding of the brain.

In this review, four interrelated topics are covered: 1) classification/regression tasks, which are often studied in the context of brain-based biomarker studies, and key DL models; 2) DL-based dynamic analysis methods, which are useful for leveraging functional information in neuroimaging data; 3) multimodal fusion methods, which are needed to leverage complementary information among the modalities; and 4) visualization and subtype

Table 1. A comparison of natural images and neuroimaging data.

	Natural images	Neuroimaging
Data set acquisition	Easy to acquire. Available samples in benchmark data sets usually number more than 1 million.	Costly to acquire. Available samples in benchmark data sets are usually less than 10,000 and often less than 10^3.
Feature characteristic	Features are usually 2D images or videos. Images under natural light. Noise distributions are mostly Gaussian.	Features are usually 3D volumes or 4D time sequences. Mostly consist of radiological images. Noise distributions vary, such as Rician noise in MRI and quantum noise in CT.
Data set labeling	Solid and intuitive ground truth. Easy to label, specific skills are not necessary.	No solid ground truth. Difficult to label, specialized skills are necessary.
DL training	Pretrained models can be used, such as VGG (https://keras.io/api/applications/).	Few pretrained models can be used. Models typically must be trained from random initialization.
DL interpretation	The effectiveness of interpretation results is intuitive.	The effectiveness of interpretation results needs to be further validated.

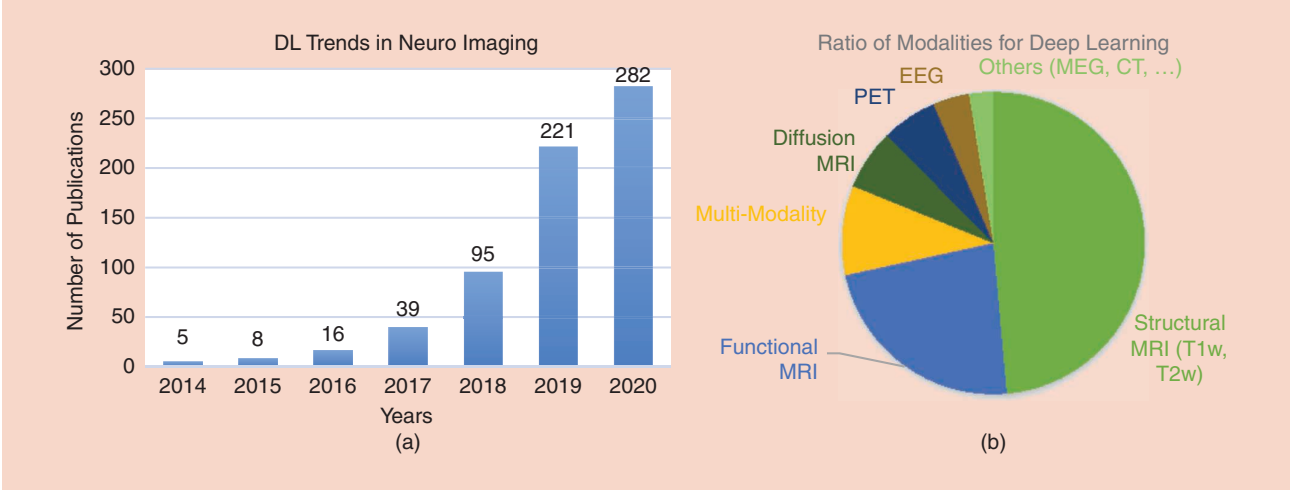


FIGURE 1. (a) DL publication trends in neuroimaging. PubMed: "deep learning" & ("brain imaging" or neuroimaging). (b) The prevalence of different neuroimaging modalities in DL studies. T1w: T1 weighted image; T2w: T2 weighted image; PET: positron emission tomography.

discovery, which is crucial for moving to clinical applications and providing clues regarding the underlying biological mechanisms.

Neuroimaging studies use a variety of techniques to study the structure and function of the nervous system, revealing the relationship between brain mechanism and behavior (symptoms). Most of the analysis approaches in neuroimaging fall into two broad categories: classification or regression. In the section "DL for Neuroimaging Classification and Regression," besides introducing the core concepts behind neural networks and DL, we summarize the architecture of the six most popular DL models and their neuroimaging application scenarios. In the section "DL for the Analysis of Dynamic Activity and Connectivity in Neuroimaging Data," we review DL methods that can leverage the information of temporal fluctuations in neuroimaging. The surveyed study of brain dynamics shows great potential for decoding brain activity and functional connectivity in various contexts, providing a window into interactions among circuits, networks, and regions, and their link to behavior in both time and space. In the section "DL for Multimodal Fusion," we review DL-based multimodal fusion models leveraging nonlinear complementary information from various modalities including brain structure, function, network connectivity, and behavior. Especially in the context of data that have mismatched dimensionality (e.g., brain structure and brain function), the flexibility of DL models is of notably great importance. In the section "Visualization and Subtype Discovery," topics related to visualization and spectrum discovery are covered. While DL is often treated as a black box, its use in studying the brain hinges on approaches to visualize and interpret important features, which can help us explore heterogeneity across healthy individuals or mental disorders. At the end of each section, we also highlight some of the promises and challenges of DL within each given category. Finally, we discuss a major challenge involving all four of the interrelated topics highlighted in this article (Figure 2): the promise of DL to accomplish important practical goals and to facilitate translational research to clinical practice.

DL for neuroimaging classification and regression

Classification and regression are two widely studied supervised-learning tasks. The difference between classification and regression tasks lies in whether the target variable is continuous or discrete. In broad terms, the core aim of both tasks is to map x (neuroimaging data) to y (e.g., diagnosis, treatment response, and behavior). Compared with natural images, neuroimaging is more complex, usually with a higher dimensionality (often above 10^4 voxels), smaller sample sizes (fewer than 10^4 samples), multiple data modalities (e.g., MRI and CT), and often lacking a solid ground truth. Even though neuroimaging data are highly diverse, two broad categories can be distinguished: structural imaging and functional imaging (Figure 2).

Structural neuroimaging data, such as those from structural MRI (sMRI) or diffusion MRI (dMRI), reflect voxel tissue density/volume or structural connectivity. The main purpose of structural studies is to reveal the anatomical relationships in the brain,

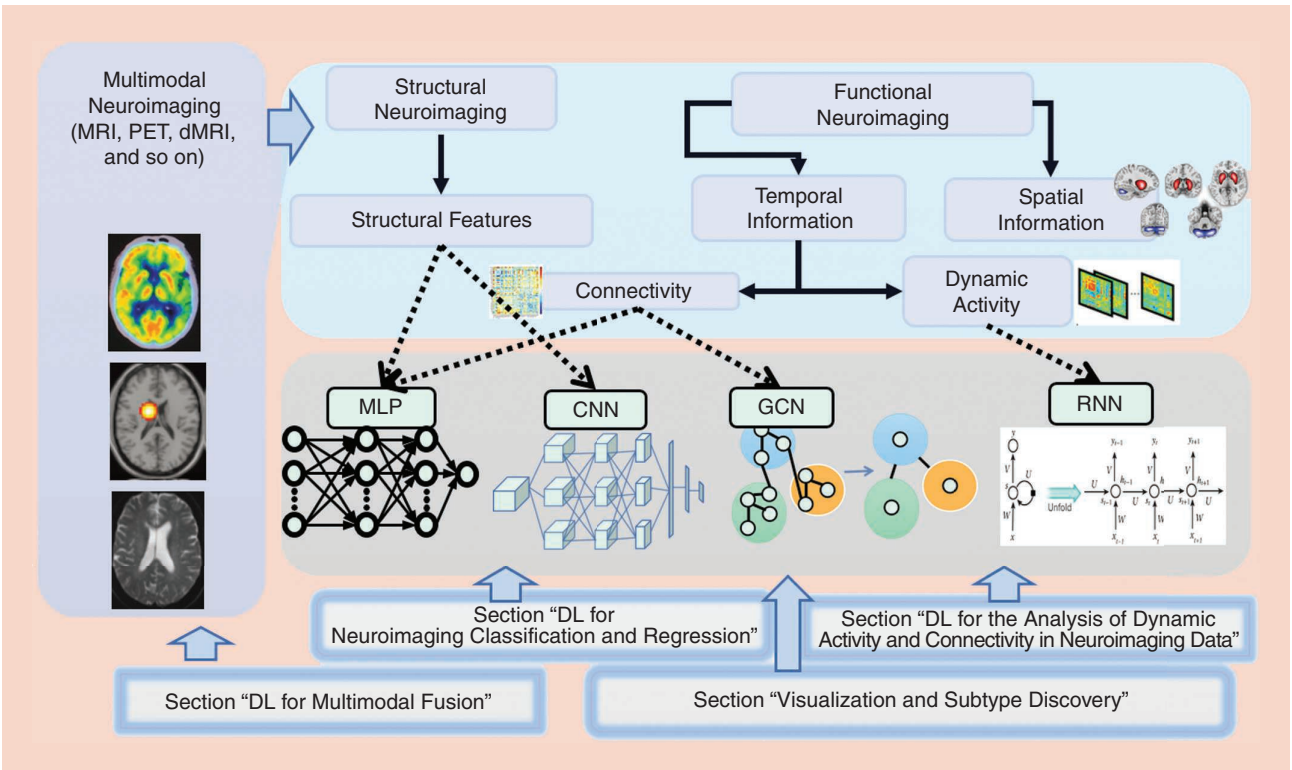


FIGURE 2. The four interrelated topics covered in this review article. Neuroimaging data are often collected from multiple modalities, being preprocessed differently to extract the input features for DL. In the gray panel, multiple DL modules are listed and linked with their applicable features. PET: positron emission tomography; dMRI: diffusion MRI; MLP: multilayer perceptron; CNN: convolutional neural network; GCN: graph convolutional network; RNN: recurrent neural network.

which can in turn be used for prediction. Functional neuroimaging data focus on dynamic changes in the activity or connectivity of the brain. Because of the high dimensionality and low signal-to-noise ratio in neuroimaging data such as MRI, efficient feature processing is important for reducing redundancy before modeling. For example, fMRI time courses are often dimensionality reduced using atlas-based or data-driven approaches, such as ICA. The resulting temporal signatures are then used for studying temporal dependence, such as functional network connectivity (FNC) or dynamic FNC (dFNC). In a typical deep learning system, there may be hundreds of millions of adjustable weights, and such models require a large number of samples for training. Properly balancing a DL model's complexity with the available sample size is vital for neuroimaging. Here, we summarize the fundamental mechanisms of the popular DL models and provide recommendations regarding their corresponding neuroimaging applications.

Multilayer perceptron models

A multilayer perceptron (MLP) model trained by simple statistic gradient descent was the first proposed solution for replacing engineered features with trainable multilayers [3]. The hidden layers can be regarded as distorting the input in a nonlinear way so that categories become separable by the last layer. This deep neural network can theoretically fit any mapping relationship. However, fully connected layers may cause redundancy of trainable parameters and overfitting. While regularization rules and dropout can remediate the overfitting problem, MLP is most suitable for low-dimensional and less redundant input, such as FNC vectors [4]. In addition, because of its flexibility, MLP is often used as a backbone for more complex DL models [e.g., a generative adversarial network (GAN)] for classification [5].

Convolutional neural networks and graph convolutional networks

Convolutional neural networks (CNNs) are now the dominant approach for almost all recognition and detection tasks. They are designed to process data that come in the form of multiple arrays, such as natural signals or images. The core elements of a CNN that take advantage of the properties of natural signals are local connections, shared weights, pooling, and the use of deep layers. The two operations that distinguish CNNs from other DL models are convolution and pooling. The role of convolutional layers is to detect local conjunctions of features from the previous layer; the role of the pooling layer is to merge semantically similar features into one. Based on the convolutional operation, the CNN input ideally consists of a highly correlated local group of values, with the local statistics of data invariant to location. Therefore, the CNN is well suited to process 2D or 3D T1 images by leveraging the spatial information to improve performance. For example, recent work has shown that leveraging the 3D structure of neuroimaging data via a CNN has substantial improvement over SML models [6]. Despite the great successes of CNNs, the non-Euclidean characteristic of graph features such as those obtained from FNC makes the general convolution and filtering not as well defined as on natural images. Similarly, a graph convolutional network (GCN) is a type of neural network architecture that can capture the graph structure and aggregate node information from the neighborhoods in a convolutional fashion with fewer learnable parameters. GCNs are useful in medical or biochemical applications with graph data, such as FNC.

Recurrent neural networks

Recurrent neural networks (RNNs) process an input sequence one element at a time, maintaining in their hidden units a "state vector" that implicitly contains information about the history of all of the past elements of the sequence. It models the following 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 nonlinear and potentially complicated, and accepts optional input $u(t)$. The long short-term memory network (LSTM) and gated recurrent unit (GRU) are two variants of RNNs that use special hidden units for remembering inputs for a longer time. Compared to classical linear machine learning models, such as a hidden Markov model, an RNN models the long-term nonlinear mechanisms of the sequential data. Therefore, the RNN is suitable for solving tasks that involve sequential inputs, such as fMRI time courses [7].

GANs

A CNN/RNN model that is trained for mapping high-dimensional features to labels is best categorized as a discriminative model because it is not focused on learning the distribution of the features. A generative model that can approximate the distribution of inputs is more robust and interpretable. Just as in a quote from Richard Feynman: "What I cannot create, I do not understand," a trained GAN model can generate samples by passing random noise through MLPs.

A GAN has two agents: a generator G and a discriminator D. G has no direct access to real data; the only way it learns is through its interaction with D. D has access to both the synthetic samples and samples drawn from the stack of real data. An error signal to D is provided through the simple ground truth of knowing whether the data came from the real stack or G. The same error signal, via D, can be used to optimize G, leading it toward being able to produce fake data of better quality [8]. The GAN is not a specific model but a generative framework. All of the previously mentioned DL models, such as MLP or CNNs, can be used as the backbone of a GAN. Compared to discriminative models, GAN models are more challenging to optimize because the data distribution is more difficult to approximate than simply finding classification borders. The representation learned by GANs may be used in a variety of neuroimaging applications, including classification, neuroimaging synthesis, and multisite neuroimaging harmonization.

Attention modules

The use of an attention module was proposed to increase the representation power and improve interpretability by focusing on important brain regions and suppressing unnecessary ones, which is often combined with other DL models for interpretation, allowing the model to dynamically emphasize certain parts of the

input. As reported in [9], a weak supervised-learning-based DL consisting of a backbone network with an attention module has been applied to improve Alzheimer's disease classification performance using sMRI. Attention maps can also be helpful for discovering task-related biomarkers. For example, an attention-guided RNN model was used for explaining the fMRI features' significance when identifying schizophrenia [10]. A transformer is a promising attention model that has no recurrent networks but can remember how sequences are fed into the model and encodes the relative position of each element [11]. These positions can be added to the embedded representation (an n-dimensional vector) of each time step of fMRI time courses.

Promises and challenges

DL has achieved great success in classification and regression tasks, and with the growing availability of data, the performance will continue to improve. However, there are still some hurdles that must be overcome. The first one is the difficulty of model design. Even though some autodifferential platforms have greatly simplified the procedures of model design, various hyperparameters, such as width, depth, loss function, and optimizers are typically decided based on experience. Fundamental DL theories, standard criteria, and handbooks are needed to guide the design of DL models. Another challenge that arises quite often in neuroimaging is that of high-dimensional small-sample problems. DL models designed for 3D or 4D neuroimaging data often consist of millions of parameters that require many samples for optimization. Large-scale neuroimaging data sets are not easily acquired, and the noise distributions vary. Thus, augmentation approaches for natural images are not well suited for neuroimaging data. Multiset data integration/fusion and improved algorithms are needed to address these domain discrepancies.

DL for the analysis of dynamic activity and connectivity in neuroimaging data

Cognition, perception, and movement arise from nonlinear dynamic activity across large-scale systems of the brain. These functions are driven by latent mental processes and external tasks. The characterization of brain activity and connectivity dynamics (e.g., the chronnectome) is crucial for our understanding of brain function [12]. However, uncovering relevant transient patterns in brain function is challenging because of the lack of computational tools that can effectively capture nonlinear dynamics from high-dimensional data. Recent studies show that DL models, especially RNN-based networks, have the potential to capture whole-brain dynamic information and utilize the time-varying functional connectivity state profiles to expand our understanding of brain function and disorder [13], [14].

Modeling spatiotemporal dynamics using DL models

Conventional neuroimaging classification approaches, which use functional network connectivity or spatial maps as input features, ignore the temporal dynamic information. DL models exhibit excellent feature representation learning ability and provide a potential tool for capturing spatiotemporal information directly from the time courses. In particular, RNNs have achieved great successes in sequence modeling tasks and are now broadly used in brain dynamic analysis for brain disorder diagnosis, brain decoding, and temporally dynamic functional state translation detection. dFNC is an approach to identify time-varying patterns of connectivity from fMRI data. To capture the temporal information in dFNC, Yan et al. [15] proposed a full-bidirectional LSTM that can handle both preceding and succeeding information by using two hidden layers with opposite information flow directions and thus better characterize the "chronnectome" (see "RNN for dFNC" in Figure 3). To overcome the effect of the window size parameter when processing the data, a CNN is used to directly extract the functional connectivity. A multiscale RNN can then incorporate spatiotemporal information in the fMRI time courses in the context of a group discrimination task (e.g., a schizophrenia diagnosis) and boost predictive performance by combining the CNN and RNN [7] (see "RNN for time courses" in Figure 3). RNN-based models can also be applied to adaptively capture temporal dependencies, providing more discriminative information for brain state decoding and prediction in real time [16]. These studies show the potential of DL models for studying brain dynamic activity, and this progress will undoubtedly continue as more models are developed.

The combination of DL with conventional neuroimaging tools

To facilitate the discovery of the dynamic information in neuroimaging data, DL can be blended with well-studied data-driven machine learning approaches, such as ICA, which can also enhance the interpretability of the results. As shown in [14], Kazemivash and Calhoun proposed a novel spatiotemporal network for brain parcellation, which combined a 3D CNN with ICA and enabled the framework to explore high-dimensional (5D) brain dynamics (see "DL combined with ICA" in Figure 3). In addition, RNN-ICA [17] has been proposed to combine RNN with ICA for a sequential ICA objective, which can explicitly optimize linear generative models to model temporal dynamics and infer intrinsic networks from time-series observations (the network structure and identified spatial maps are shown in "RNN leverages ICA" in Figure 3). RNN-ICA extends the RNN frameworks to incorporate the infomax objective and can be applied to various types of data (e.g., simulated synthetic data, task-related scans, and resting-state fMRI) to identify both similar task-relatedness patterns and directed temporal connectivity.

Promises and challenges

RNN-based models take advantage of the ability to model the sequential information/dynamic functional connectivity and simulate the periodic brain status change; therefore, they can achieve an improved predictive performance compared with conventional models. However, in existing work, dynamic features are often calculated using window-based correlation, and thus the window size is a hyperparameter that affects the dFNC features. Window-based methods with a short window cannot capture long-time correlations, whereas longer windows reduce sensitivity to rapid changes; hence, it can be challenging to select the proper window

size. Recent research in natural language processing has proposed transformer models [11] that can capture sequential interdependence using the attention mechanism to provide a potential brain dynamic modeling solution. Beyond this, it can be challenging to validate the results because of the lack of a gold-standard ground truth. Considering that most existing measurements focus on comprehensive assessments rather than temporally targeted information, additional studies are needed to evaluate the reliability and reproducibility of the analyzed results.

DL for multimodal fusion

Neuroimaging data typically include multiple modalities, such as sMRI, fMRI, and dMRI, which provide multiple views for observing and analyzing the brain. To leverage the complementary representations of different modalities, multimodal fusion is consequently needed to provide a more complete understanding of brain mechanisms [18]. However, conventional nonlinear fusion models may not be sufficiently flexible to fully capture intrinsic structures and external relationships in multimodal neuroimaging. DL multimodal fusion methods, which can learn multilevel nonlinear abstract representation of the data, have outperformed conventional fusion methods in many tasks.

DL frameworks for multimodal fusion

A variety of DL models, including all those mentioned earlier, have been applied as the backbone to extract high-level features in multimodal neuroimaging fusion frameworks. The model choice depends heavily on the data structure of each modality. Despite the variety of available models, most multimodal fusion strategies fall into the following two categories: prefusion and postfusion. A prefusion strategy concatenates raw features from multiple modalities before sending them to DLs. By contrast, a postfusion strategy first employs DLs for learning feature representations of each modality and then concatenates them for subsequent tasks. Prefusion is easy to implement but has limitations when the feature dimensionality of one modality is much higher than the others or when the concatenation is infeasible because of the heterogeneity in data format. Compared to prefusion, a postfusion framework is more flexible when dealing with diverse modalities but more laborious in finding the optimal architectures and hyperparameters.

Beyond the concatenation-based postfusion, more advanced postfusion methods have been proposed by considering cross-modality relationships. Multimodal reconstruction, deep canonical correlation analysis (DCCA), and knowledge-transfer-based

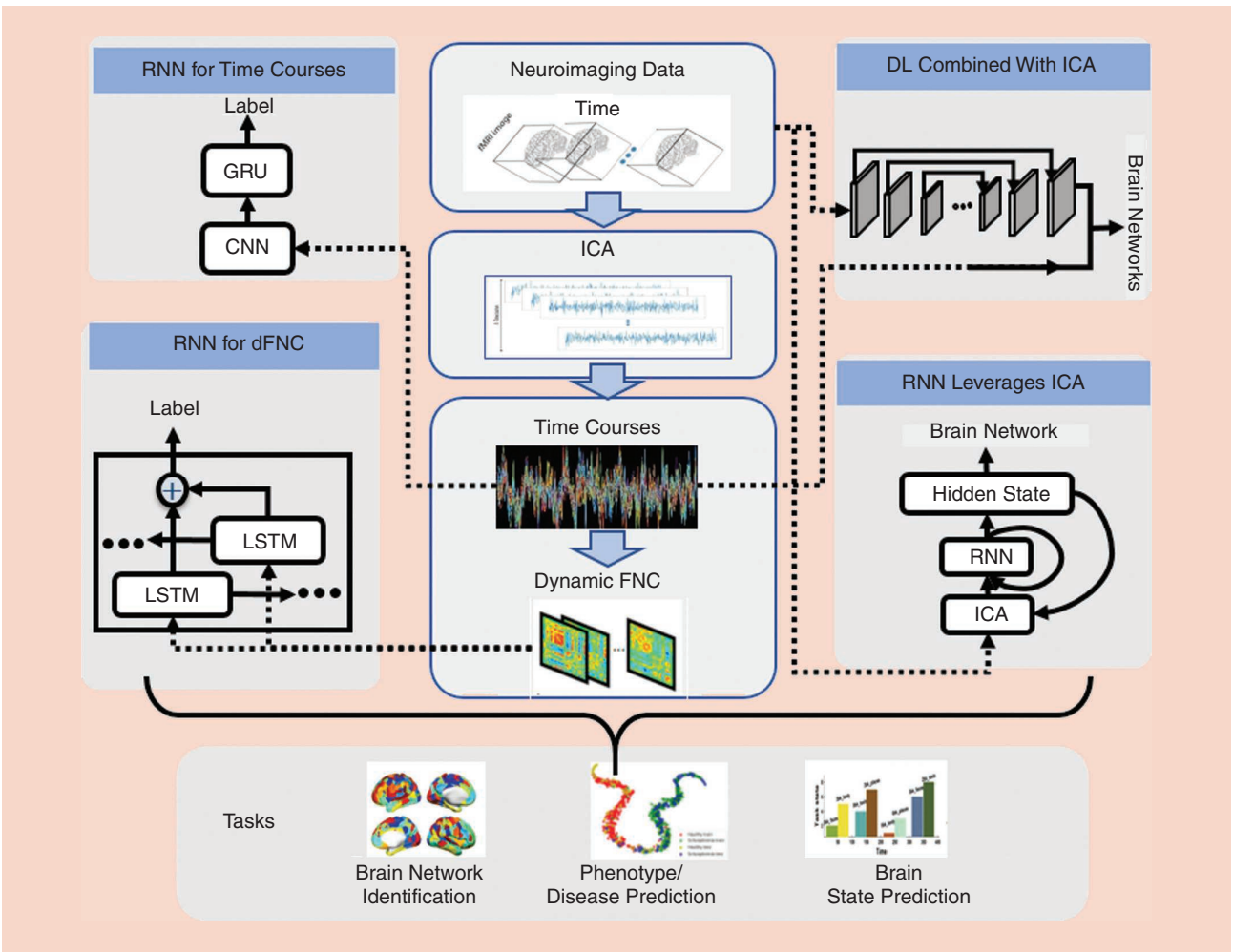


FIGURE 3. DL for analysis of dynamic activity and connectivity in neuroimaging data. The different stages of sequential features (e.g., time courses) extracted from neuroimaging data are processed using suitable DL models (e.g., RNN) to facilitate various tasks (e.g., brain network identification).

fusion are three popular multimodal fusion methods. As illustrated in Figure 4, a multimodal reconstruction method employs autoencoders (AEs) to learn optimal cross-modality representations that can best reconstruct the original data. Unlike a standard AE, multimodal reconstruction learns a representation with two encoders and then uses the shared representation for reconstruction, which is suitable for unsupervised tasks where the label is not acquired. Capturing cross-modality correlation or mutual information is another way to perform multimodal fusion. One example is DCCA, which allows two DL models to learn new representations while optimizing their correlations. The fusion performance of DCCA can be further improved by utilizing knowledge transfer, which retains the correlated features and leverages information among different modalities [19].

Multimodal fusion applications in neuroimaging

The availability of multiple neuroimaging data and the complexity of the brain have led to numerous multimodal fusion applications. For example, Venugopalan et al. compared prefusion and postfusion frameworks by integrating MRI imaging data, electronic health record data (including longitudinal information about patients and doctors), and SNP data for Alzheimer's disease identification [20]. The results showed that postfusion worked better than prefusion because of the high data heterogeneity. Some other state-of-the-art cross-modality representation methods that can better learn latently shared and distinguished relationships have also been proposed. Deep collaborative learning can incorporate labels into the DCCA method. It has been validated on resting-state fMRI and task fMRI [21], showing high performance for classifying age groups. A combination of an AE and the DCCA method was proposed to better classify schizophrenia by integrating fMRI data and SNP data [22] [Figure 5(a)]. In addition, a multimodal GCN achieved high performance in a cognitive-ability prediction task by using a manifold to regularize the multimodal GCN and considering the relationships of subjects both within and between modalities [19] [Figure 5(b)]. Plis et al. proposed a translation-based fusion model that learned the linkage between functional dynamic connectivity and static gray matter patterns computed from sMRI. The work was evaluated on multisite resting-state MRI data, also including an independent data set [23].

Promises and challenges

The development of state-of-the-art fusion methods (e.g., cross-modal representation-based methods) has shown enhanced performances over unimodal analysis within the DL framework, which facilitates the early detection or subtype classification of brain diseases from comprehensive views. However, multimodal fusion often lacks enough training samples. In addition, most approaches require modalities to be available for all data sets, resulting in samples being discarded. The choices of models and fusion strategies in existing works are usually based on intuition. Thus, a quantitative explanation of how the high-level features are extracted and how they contribute to the downstream tasks is needed. In addition, since joint features extracted from various modalities are aggregated within a unified model, the roles of each modality can be blurred. Therefore, interpretation can be even more challenging in the context of multimodal data fusion.

Visualization and subtype discovery

The flexibility of DL, including its ability to model nonlinear relationships, is vital but can make interpretation challenging. In contrast to natural images, neuroimaging studies often lack a

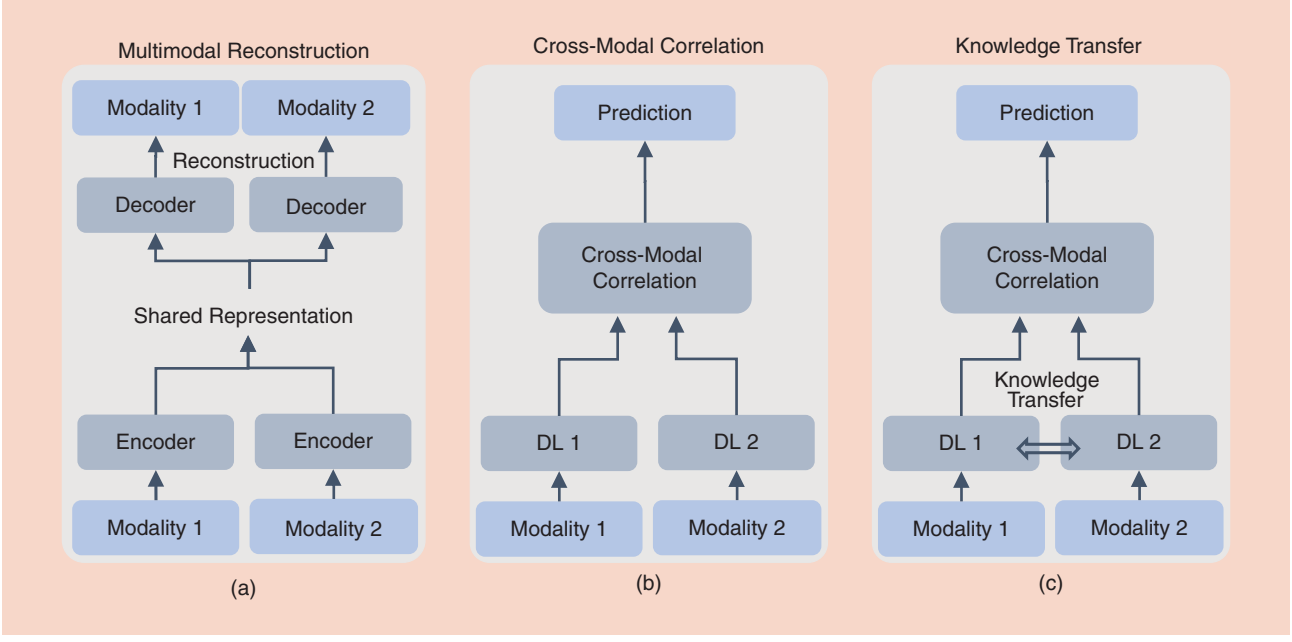
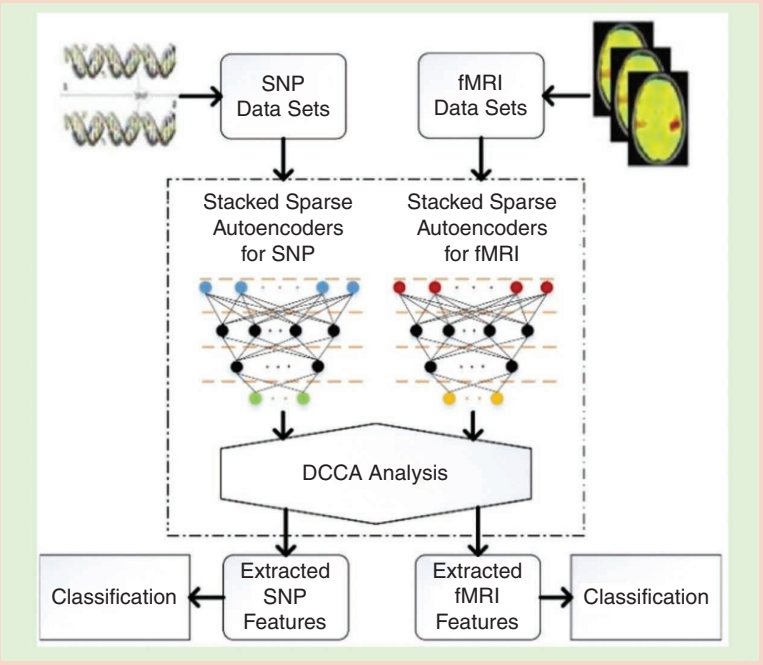
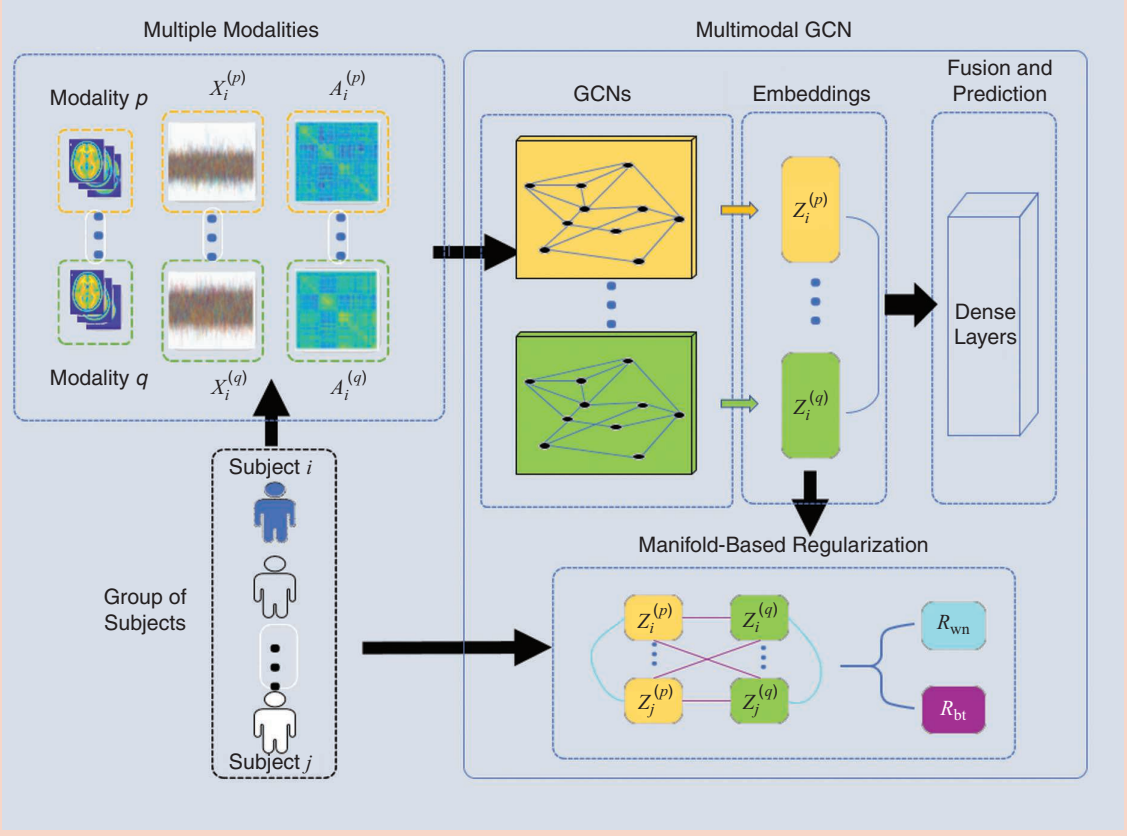


FIGURE 4. Three cross-modality representation-based multimodal fusion DL frameworks. (a) The multimodal reconstruction method employs AEs to learn optimal cross-modality representations that can best reconstruct the original data. (b) Cross-modal correlation allows DL models to learn new representations while optimizing their correlations. (c) The knowledge-transfer model further considers the manifold regularization between modalities. AE: autoencoder.



(a)



(b)

FIGURE 5. Multimodal fusion applications in neuroimaging. (a) An imaging-genetic integration work using a deep canonically correlated sparse AE for the classification of schizophrenia [22]. (b) A multimodal GCN with knowledge transfer within and between modalities [19]. The manifold regularization term fully explores the relationship between subjects, enforcing the model to learn similar embeddings for subjects with high brain structure similarity both within and between modalities.

solid ground truth, especially psychiatric neuroimaging studies. Because of this, DL visualization is a crucial way to expand our knowledge of clinical cues of brain disorders. Visualization is also used for discovering biomarkers and relationships among mental disorders.

Network visualization for biomarker discovery

Reasonable network visualization approaches should meet the following three requirements: 1) be readable and understandable to humans; 2) provide useful information about what mental or behavioral constructs are represented in particular brain pathways or regions; and 3) be based on relevant neurobiological signals, and not confounds [24]. The popular visualization approaches can be classified into four categories: interpretable local surrogates, occlusion analysis, gradient-based methods, and layer-wise relevance propagation [25] [Figure 6(a)].

Interpretable local surrogates produce explanations of a DL by locally approximating it 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 interpreted. Local interpretable model-agnostic explanation (LIME) and Shapley additive explanations (SHAP) are two classical implementations of local surrogates. Lombardi et al. embedded SHAP and LIME to explain the outcomes of DL models by determining the contribution of each brain morphological descriptor to the final predicted age of each subject and investigating the reliability of the two methods. The SHAP approach was found to provide more reliable explanations for morphological aging mechanisms [26]. Occlusion analysis is a widely used architecture-independent method in which a particular type of perturbation analysis repeatedly tests the effects on the neural network's output when occluding patches or individual features in the input features. A heatmap is built from these scores, highlighting locations where the occlusion has caused the strongest effect of the function. Occlusion analysis has been applied to CNN- and RNN-based models for measuring the contribution of each brain region in classification tasks [7].

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 versus controls has been performed using a specific gradient-based implementation [27]. However, gradient-based methods are often computationally expensive, especially when making the integration procedure

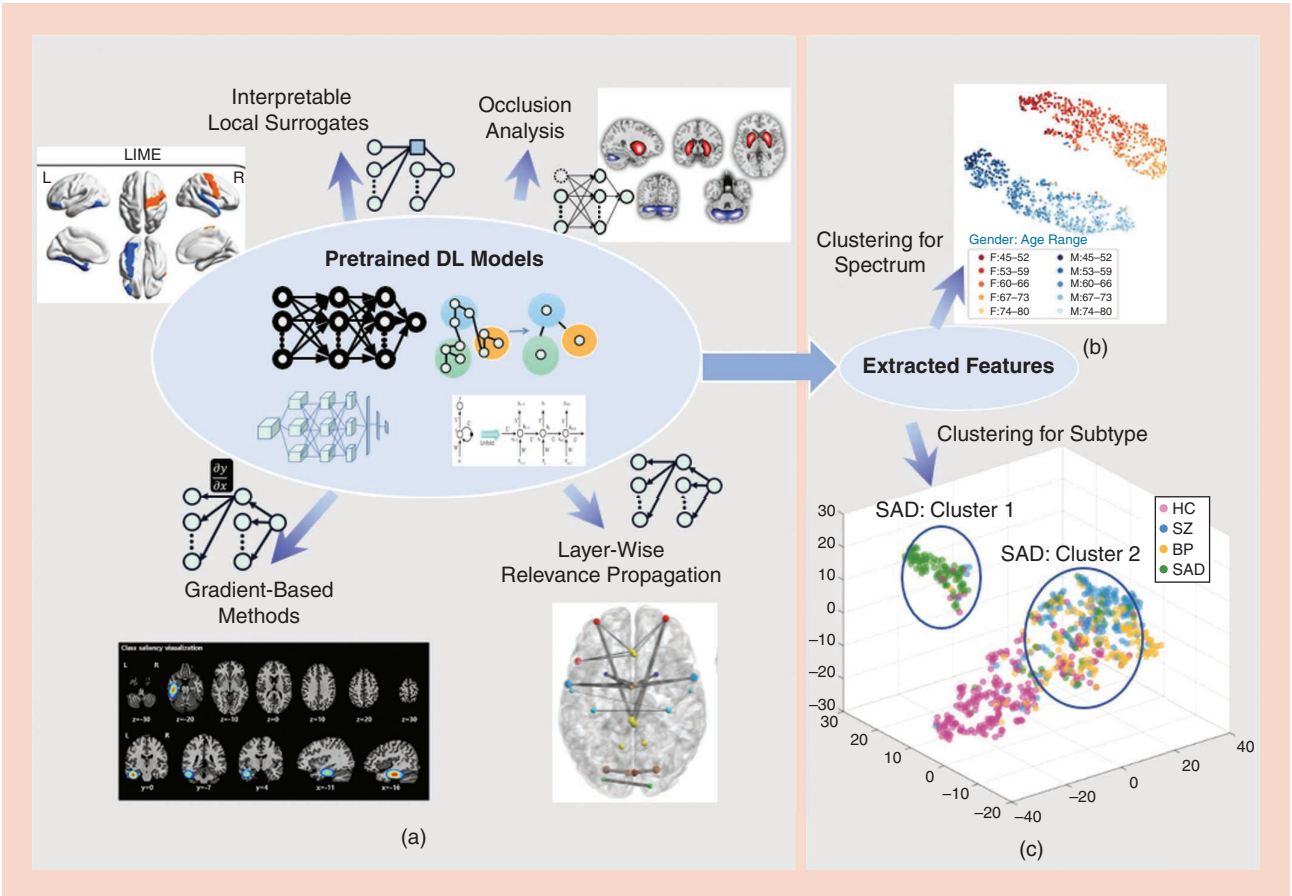


FIGURE 6. DL interpretation and visualization in neuroimaging. (a) Four DL interpretation approaches in neuroimaging. By analyzing the pretrained DL models, discriminative features are visualized to provide insight into their use as potential biomarkers. (b) The highly abstracted features extracted by DL can be further clustered for spectrum visualization. (c) The subtype of mental disorders can be discovered using extracted features. SAD: schizoaffective disorder; SZ: schizophrenia; BP: bipolar disorder; HC: healthy control.

accurate. Layer-wise relevance propagation makes explicit use of the layered structure of the neural network and operates iteratively to produce an explanation. Layer-wise relevance propagation analysis is performed at the level of single input samples, enabling an analysis on several levels of data granularity, from the level of the group down to the level of single subjects, trails, and time points [28].

Spectrum and subtype discovery using DL framework

Psychiatric disorders are often diagnosed based on symptoms rather than biological data. There is also often considerable overlap among different types of psychiatric disorders, which makes accurate diagnosis challenging. Examining the neurobiology of the psychotic-affective spectrum may greatly advance the biological determination of psychiatric diagnosis, which is critical for the development of more effective treatments [29]. DL can jointly optimize feature embedding and classification hyperplanes using the error backpropagation method. As shown in Figure 6(b), Abrol et al. projected the learned DL embedding onto a 2D plane using *t*-distributed stochastic neighbor embedding (tSNE) for the entire range of training samples and color-coded the 2D projection spectrum by the class labels. They found separate gender clusters, ordered in increasing age from one end of the spectrum to the other [6]. Similar results have also been obtained when using DL to discriminate Huntington's disease based on MRI [1]. The results indicate that DL encodes more robust nonlinear discriminative neuroimaging representations than conventional machine learning.

Subtype discovery is crucial to move toward precise medicine, such as individualized therapy, but it is also challenging, especially when the signal-to-noise ratio is low. Under such circumstances, clustering models are likely to be misled by confounds such as age, gender, or site effects. To overcome this problem, DL can be used to map the neuroimaging data into a subspace in which the subtypes can be clustered. The supervised classification module can be first trained using a supervised way to map the original fMRI features to a subspace where the differences among psychiatric disorders are more distinctive. Then high-level representations of the original features are submitted to a tSNE clustering model for visualizing the group differences among disorders [30] [Figure 6(c)].

Promises and challenges

Interpreting ML models in neuroimaging is intrinsically an open-ended process. The developing DL interpretation approaches show promise for providing insights into new mechanisms of brain activity and biomarkers of brain disorders. Unlike natural image data sets, which contain millions of accurately labeled training samples, the ground truth is usually not clear in neuroimaging studies, and the cost of the incorrect interpretation is high. For instance, even a sophisticated psychiatrist cannot tell the differences between a patient with depression and a healthy control based merely on fMRI. Because of this, DL interpretation methods that might work well in the natural imaging field cannot easily be applied to the neuroimaging field because it is difficult to validate the results. In addition, different explainability approaches may not always obtain consistent results. The effectiveness of the results should be validated using various invasive techniques (e.g., brain stimulation).

Future directions: From the lab to clinical practice

The strength of DL models is that they can implement complicated and, in principle, arbitrary, predictor–response mappings efficiently. This power comes with some costs, including the requirement of a large number of training samples, complicated model architectures, and difficulty in model interpretation. Despite promising results in neuroimaging analysis, few algorithms have reached clinical implementation, challenging the balance between hope and hype for these techniques. The real clinical value of machine learning methods and their associated biomarkers will likely come from our ability to detect subtle differences in imaging signatures before the disease is clinically diagnosed, to refine clinical categories according to imaging phenotypes of clinical relevance, or to inform treatment.

Minimizing the model design and model fine-tuning burden

The widespread success of DL methods has created a need for architecture engineering, where data scientists are tasked with manually designing increasingly complex neural architectures. The neural architecture search (NAS) technique has emerged, which seeks to automatically select, compose, and parameterize DL models to achieve optimal performance on a given data set and task. NAS methods are best categorized by three factors: search space, search strategy, and performance estimation strategy. The search space refers to the potential neural architectures that can be represented by the NAS algorithm, and the search strategy refers to how this space is explored. The performance estimation strategy refers to how the NAS algorithm evaluates a given architecture's performance on some tasks given some training data set. NAS is an important but relatively new field in neuroimaging.

Privacy protection in multisite collaboration

Multisite collaboration is necessary to gather more data for DL training. Instead of transferring data directly to a centralized data warehouse for building machine learning models, federated (or decentralized) learning enables multiple sites to collaboratively learn a shared classification/prediction model while keeping the training data at each local site. As shown at http://coinstac.trendscenter.org, a local site can download the current DL model and improve it by learning from data on its site and then summarize the changes as a minor focused update. Such an update can then be uploaded to the cloud, providing a scalable option for accessing more data via multisite collaboration and privacy protection.

Interpretation results and clinical validation

A concise interpretation result should not only be relatively consistent when using different interpretation methods but also generalizable to other data sets or tasks [24]. Going forward, it will be imperative to bring in more converging evidence from related literatures and invasive studies (e.g., transcranial magnetic stimulation or electroconvulsive therapy) with different modalities

and multiple species to better understand the model's neurobiological meaning.

Conclusions

DL, which allows computational models consisting of multiple processing layers to learn representations of data with multiple levels of abstraction, is a promising method and has been making breakthroughs in the neuroimaging field. In this work, we systematically review the basic mechanisms of DL in neuroimaging and highlight some key findings, including the following. 1) DL is able to outperform SML in large-scale neuroimaging classification and regression tasks when using rich features. 2) When incorporated with dynamic analysis, DL shows strength in capturing time-varying information and can improve the sensitivity and specificity. 3) By leveraging complementary, multifaceted information, multimodal fusion combined with DL is more efficient and flexible than traditional methods. 4) Complex nonlinear relationships in neuroimaging can be captured by DL to identify novel disease subtypes, facilitating biomarker discovery.

The development of imaging techniques and multisite collaboration and data sharing is producing the additional high-quality neuroimaging data needed to fuel DL to uncover key brain mechanisms. Combining DL interpretation with invasive methods will lead to more reliable biomarkers with the potential for clinical value. In conclusion, DL opens a window for exploring brain mechanisms through the lens of many types of neuroimaging features. As a result, the field is rapidly moving toward more refined and biologically based diagnoses as well as precise clinical applications.

Acknowledgments

This work was supported by the National Institutes of Health (grants R01EB005846, R01MH117107, R01GM109068, R01MH104680, R01MH107354, and R56MH124925), the National Science Foundation (1539067 and 2112455), and the Natural Science Foundation of China (82022035, 61773380, and 12090021). Weizheng Yan, Gang Qu, and Wenxing Hu contributed equally to this work. The corresponding authors are Vince D. Calhoun and Jing Sui.

Authors

Weizheng Yan (wyan3@gsu.edu) received his Ph.D. degree in pattern recognition and intelligence systems from the Institution of Automation, Chinese Academy of Sciences in 2020. He is now a postdoctoral research associate at the Tri-Institutional Center for Translational Research in Neuroimaging and Data Science, Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, Georgia, 30303, USA. His research interests include designing deep learning algorithms for neuroimaging analysis.

Gang Qu (gqu1@tulane.edu) received his M.S. degree in biomedical engineering from the University of Florida, Gainesville, Florida, in 2018. He is currently a Ph.D. student in biomedical engineering at Tulane University, New Orleans, Louisiana, 70118, USA. His research interests include graph theory, machine learning, and deep learning and their applications to the integration of multimodal biomedical data.

Wenxing Hu (whu@tulane.edu) received his B.Sc. degree in applied mathematics from Xi'an Jiaotong University, China, in 2011. He is currently a Ph.D. student in biomedical engineering at Tulane University, New Orleans, Louisiana, 70118, USA. His research interests include dimension reduction, correlation analysis, and multi-omics data integration.

Anees Abrol (aabrol@gsu.edu) received his Ph.D. degree in electrical engineering from The University of New Mexico in 2018. He is currently a research scientist at the Tri-Institutional Center for Translational Research in Neuroimaging and Data Science, Georgia State University, Georgia Institute of Technology, and Emory University, Atlanta, Georgia, 30302-3965, USA. His research involves developing advanced machine learning- and signal processing-based frameworks to understand complex interactions in multimodal brain imaging data and explain the underlying structural and functional brain changes in neuropsychological diseases.

Biao Cai (bcai1@tulane.edu) received his M.S. degree in biomedical engineering from Tianjin University, China. He is currently a Ph.D. candidate in biomedical engineering at Tulane University, New Orleans, Louisiana, 70118, USA. His research interests include machine learning, deep learning, and their applications in time-varying analysis and individual identification of functional magnetic resonance imaging data. He is a Member of IEEE.

Chen Qiao (qiaochen@xjtu.edu.cn) received her Ph.D. degree in applied mathematics in 2009 from Xi'an Jiaotong University, China. In 2014–2015, she was a postdoctoral researcher in the Department of Biomedical Engineering, Tulane University, New Orleans, Louisiana. In 2019, she was a research fellow in the School of Computer Science and Engineering, Nanyang Technological University. Currently, she is a full professor in the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, 710049, China, and the director of the Brain Science Laboratory of Xi'an Jiaotong University SuZhou Academy. Her current research interests include the mathematical foundation of information technology, artificial intelligence, and neuroimaging.

Sergey M. Plis (splis@gsu.edu) received his Ph.D. degree in computer science from the University of New Mexico, Albuquerque, New Mexico, in 2007. He is currently an associate professor of computer science with the Georgia State University and the director of the machine learning core with the Tri-Institutional Center for Translational Research in Neuroimaging and Data Science, Atlanta, Georgia, 30303, USA. His research interests include developing novel approaches to analyze large-scale data sets in multimodal brain imaging and other domains.

Yu-Ping Wang (wyp@tulane.edu) received his Ph.D. degree in communications and electronic systems from Xi'an Jiaotong University, Xi'an, China, in 1996. He is a professor of biomedical engineering and biostatistics and bioinformatics with the School of Science and Engineering, Tulane University, New Orleans, Louisiana, 70118, USA, and with the School of Public Health and Tropical Medicine, New Orleans, Louisiana. His

research interests include computer vision, signal processing, and machine learning with applications to biomedical imaging and bioinformatics. He is a Senior Member of IEEE.

Jing Sui (jsui@bnu.edu.cn) received her Ph.D. degree in optical engineering with honors from Beijing Institute of Technology in 2007. She worked at the Mind Research Network, New Mexico, USA, as a postdoctoral fellow and was promoted to research scientist in 2010 and assistant professor of translational neuroscience in 2012. She is currently a full professor at the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, 100875, China. Her research interests include machine learning in neuroimaging, multimodal brain imaging data fusion, pattern recognition, and their applications in mental illnesses. She is a Senior Member of IEEE.

Vince D. Calhoun (vcalhoun@gsu.edu) received his Ph.D. degree in electrical engineering from the University of Maryland Baltimore County. He directs the Tri-Institutional Center for Translational Research in Neuroimaging and Data Science with appointments at Georgia State, Georgia Tech, and Emory, Atlanta, Georgia, 30302-3965, USA. He is the author of over 900 peer-reviewed papers. He develops flexible methods to analyze neuroimaging data. He is a Fellow of IEEE and a fellow of the American Association for the Advancement of Science, American Institute for Medical and Biological Engineering, American College of Neuropsychopharmacology, Organization for Human Brain Mapping, and the Institute for International Society for Magnetic Resonance in Medicine. He serves on the IEEE Bio Imaging and Signal Processing Technical Committee (TC), the IEEE Data Science Initiative Steering Committee, and the IEEE Brain TC.

References

[1] S. M. Plis, D. R. Hjelm, R. Salakhutdinov, E. A. Allen, H. J. Bockholt, J. D. Long, H. J. Johnson, J. S. Paulsen *et al*., "Deep learning for neuroimaging: A validation study," *Front. Neurosci.*, vol. 8, p. 229, 2014, doi: 10.3389/fnins.2014.00229.

[2] V. D. Calhoun, J. Liu, and T. Adalı, "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data," *Neuroimage*, vol. 45, no. 1, Supplement 1, pp. S163–S172, 2009, doi: 10.1016/j.neuroimage.2008.10.057.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 28, 2015, doi: 10.1038/nature14539.

[4] 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, pp. 127–146, Jan. 1, 2015, doi: 10.1016/j.neuroimage.2015.05.018.

[5] J. Zhao, J. Huang, D. Zhi, W. Yan, X. Ma, X. Yang, X. Li, Q. Ke *et al*., "Functional network connectivity (FNC)-based generative adversarial network (GAN) and its applications in classification of mental disorders," *J. Neurosci. Methods*, vol. 341, p. 108756, Jul. 15, 2020, doi: 10.1016/j.jneumeth.2020.108756.

[6] A. Abrol, Z. Fu, M. Salman, R. Silva, Y. Du, S. Plis, and V. Calhoun, "Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning," *Nature Commun.*, vol. 12, no. 1, p. 353, 2021, doi: 10.1038/s41467-020-20655-6.

[7] W. Yan, V. Calhoun, M. Song, Y. Cui, H. Yan, S. Liu, L. Fan, N. Zuo *et al*., "Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site FMRI data," *EBioMedicine*, vol. 47, pp. 543–552, Sept. 2019, doi: 10.1016/j.ebiom.2019.08.023.

[8] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, 2018, doi: 10.1109/MSP.2017.2765202.

[9] S. Liang and Y. Gu, "Computer-aided diagnosis of Alzheimer's disease through weak supervision deep learning framework with attention mechanism," *Sensors*, vol. 21, no. 1, p. 220, 2021, doi: 10.3390/s21010220.

[10] M. Zhao, W. Yan, R. Xu, D. Zhi, R. Jiang, T. Jiang, V. D. Calhoun, and J. Sui, "An attention-based hybrid deep learning framework integrating temporal coherence and dynamics for discriminating schizophrenia," in *Proc. IEEE 18th Int. Symp. Biomed. Imaging (ISBI)*, 2021, pp. 118–121, doi: 10.1109/ISBI48211.2021.9433919.

[11] A. Vaswani, N. Shaneer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Conf. Neural Inf. Process.*, 2017, pp. 5998–6008.

[12] V. D. Calhoun, R. Miller, G. Pearlson, and T. Adalı, "The chronnectome: Time-varying connectivity networks as the next frontier in fMRI data discovery," *Neuron*, vol. 84, no. 2, pp. 262–274, 2014, doi: 10.1016/j.neuron.2014.10.015.

[13] 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, doi: 10.3390/a14030075.

[14] 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," 2021, bioRxiv 2021.04.22.440936.

[15] 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, Sept. 2018, doi: 10.1007/978-3-030-00931-1_29.

[16] H. Li and Y. Fan, "Interpretable, highly accurate brain decoding of subtly distinct brain states from functional MRI using intrinsic functional networks and long short-term memory recurrent neural networks," *Neuroimage*, vol. 202, p. 116059, Nov. 2019, doi: 10.1016/j.neuroimage.2019.116059.

[17] R. D. Hjelm, E. Damaraju, K. Cho, H. Laufs, S. M. Plis, and V. D. Calhoun, "Spatio-temporal dynamics of intrinsic networks in functional magnetic imaging data using recurrent neural networks," *Front. Neurosci.*, vol. 12, p. 600, Sep. 2018, doi: 10.3389/fnins.2018.00600.

[18] J. Sui, S. Qi, T. G. M. van Erp, J. Bustillo, R. Jiang, D. Lin, J. A. Turner, E. Damaraju *et al.*, "Multimodal neuromarkers in schizophrenia via cognition-guided MRI fusion," *Nature Commun.*, vol. 9, no. 1, p. 3028, 2018, doi: 10.1038/s41467-018-05432-w.

[19] G. Qu, L. Xiao, W. Hu, K. Zhang, V. D. Calhoun, Y.-P. Wang, "Ensemble manifold regularized multi-modal graph convolutional network for cognitive ability prediction," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 12, pp. 3564–3573, 2021, doi: 10.1109/TBME.2021.3077875.

[20] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of Alzheimer's disease stage," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, 2021, doi: 10.1038/s41598-020-74399-w.

[21] W. Hu, B. Cai, A. Zhang, V. D. Calhoun, and Y.-P. Wang, "Deep collaborative learning with application to the study of multimodal brain development," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 12, pp. 3346–3359, 2019, doi: 10.1109/TBME.2019.2904301.

[22] G. Li, D. Han, C. Wang, W. Hu, V. D. Calhoun, and Y.-P. Wang, "Application of deep canonically correlated sparse autoencoder for the classification of schizophrenia," *Comput. Methods Programs Biomed.*, vol. 183, p. 105073, Jan. 2020, doi: 10.1016/j.cmpb.2019.105073.

[23] S. M. Plis, M. F. Amin, A. Chekroud, D. Hjelm, E. Damaraju, H. J. Lee, J. R. Bustillo, K. H. Cho, G. D. Pearlson, "Reading the (functional) writing on the (structural) wall: Multimodal fusion of brain structure and function via a deep neural network-based translation approach reveals novel impairments in schizophrenia," *NeuroImage*, vol. 181, pp. 734–747, Nov. 2018, doi: 10.1016/j.neuroimage.2018.07.047.

[24] L. Kohoutová, J. Heo, S. Cha, S. Lee, T. Moon, T. D. Wager, and C.-W. Woo, "Toward a unified framework for interpreting machine-learning models in neuroimaging," *Nat. Protocols*, vol. 15, no. 4, pp. 1399–1435, 2020, doi: 10.1038/s41596-019-0289-5.

[25] 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," *Proc. IEEE*, vol. 109, no. 3, pp. 247–278, 2021, doi: 10.1109/JPROC.2021.3060483.

[26] A. Lombardi, D. Diacono, N. Amoroso, A. Monaco, J. M. R. S. Tavares, R. Bellotti, and S. Tangaro, "Explainable deep learning for personalized age prediction with brain morphology," *Front. Neurosci., Original Res.*, vol. 15, no. 578, 2021, doi: 10.3389/fnins.2021.674055.

[27] K. Oh, W. Kim, G. Shen, Y. Piao, N.-I. Kang, I.-S. Oh, Y. C. Chung, "Classification of schizophrenia and normal controls using 3D convolutional neural network and outcome visualization," *Schizophrenia Res.*, vol. 212, pp. 186–195, Oct. 2019, doi: 10.1016/j.schres.2019.07.034.

[28] 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.

[29] H. Rokham, G. Pearlson, A. Abrol, H. Falakshahi, S. Plis, and V. D. Calhoun, "Addressing inaccurate nosology in mental health: A multilabel data cleansing approach for detecting label noise from structural magnetic resonance imaging data in mood and psychosis disorders," *Biol. Psychiatr.: Cogn. Neurosci. Neuroimag.*, vol. 5, no. 8, pp. 819–832, 2020, doi: 10.1016/j.bpsc.2020.05.008.

[30] 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 Res.*, vol. 2021, doi: 10.1016/j.schres.2021.02.007.

SP