Deep Neural Generative Model of Functional MRI Images for Psychiatric Disorder Diagnosis

Takashi Matsubara, Member, IEEE, Tetsuo Tashiro, Nonmember, and Kuniaki Uehara, Nonmember

Abstract—Accurate diagnosis of psychiatric disorders plays a critical role in improving quality of life for patients and potentially supports the development of new treatments. Many studies have been conducted on machine learning techniques that seek brain imaging data for specific biomarkers of disorders. These studies have encountered the following dilemma: An end-to-end classification overfits to a small number of high-dimensional samples but unsupervised feature-extraction has the risk of extracting a signal of no interest. In addition, such studies often provided only diagnoses for patients without presenting the reasons for these diagnoses. This study proposed a deep neural generative model of resting-state functional magnetic resonance imaging (fMRI) data. The proposed model is conditioned by the assumption of the subject's state and estimates the posterior probability of the subject's state given the imaging data, using Bayes’ rule. This study applied the proposed model to diagnose schizophrenia and bipolar disorders. Diagnosis accuracy was improved by a large margin over competitive approaches, namely support vector machine, logistic regression, and multilayer perceptron. This study proposed a deep neural network that is expected to belong to one of the classes. Each scan image visualizes brain regions largely related to the disorders, thus motivating further biological investigation.

Index Terms—deep learning, generative model, functional magnetic resonance imaging, mental-disorder diagnosis, schizophrenia, bipolar disorder

I. INTRODUCTION

ACCURATE diagnosis of neurological and psychiatric disorders plays a critical role in improving quality of life for patients; it provides an opportunity for appropriate treatment and prevention of further disease progression. Moreover, it potentially enables the effectiveness of treatments to be evaluated and supports development of new treatments. With advances in brain imaging techniques such as (functional) magnetic resonance imaging (MRI) and positron emission tomography (PET), many studies have attempted to find specific biomarkers of neurological and psychiatric disorders in brain images using machine learning techniques [2], e.g., for schizophrenia [3], [4], Alzheimer’s disease (AD) [5]–[9], and others [10]–[14]. Resting-state fMRI (rs-fMRI) has received considerable attention [7], [11], [13]. This approach visualizes brain regions in subjects at rest, that is, it does not require subjects to perform tasks and to receive stimuli, which eliminates potential confounders, e.g., individual task-skills [15].

Although neuroimaging datasets continue to increase in size [1], each dataset contains only a small number of high-dimensional samples compared to datasets for other machine-learning tasks. Unsophisticated application of machine-learning techniques is prone to overfitting to training samples and failing to generalize to unknown samples. Hence, many existing techniques are composed of unsupervised feature-extraction and supervised classification [3]–[7], [10]–[12]. These approaches first identify low-dimensional dominant patterns in the high-dimensional samples and extract the former as features using unsupervised dimension-reduction methods such as principal component analysis (PCA) [6], [10] and independent component analysis (ICA) [3], [4], [11]. Then, disorders are diagnosed based on the extracted features using supervised classifiers such as a support vector machine (SVM) [8], [10]. Unsupervised feature-extractors are considered to reduce the risk of overfitting. However, they inevitably risk extracting factors unrelated to the disorder, rather than extracting disorder-related brain activity [16].

In contrast, artificial neural networks with deep architectures (deep neural networks; DNNs) are attracting attention in the machine-learning field (see [17], [18] for a review). They have the ability to approximate arbitrary functions and learn high-level features from a given dataset automatically, and thereby improve performance in classification and regression tasks related to images, speech, natural language, and more besides. Variants of DNNs have been employed for neuroimaging datasets: a multilayer perceptron (MLP) as a supervised classifier [4]–[6], [11] and an autoencoder (AE) as an unsupervised feature-extractor [4], [5], [7], [12]. These approaches share common difficulties with the aforementioned techniques but the former are uniquely characterized by their modifiable structures: The AE can be extended to a deep neural generative model (DGM), which implements relationships between multiple factors (e.g., fMRI images, class labels, imposed tasks, and stimuli) in its network structure [19]–[23]. The DGM with class labels is no longer just an unsupervised feature-extractor but is a generative model of the joint distribution of data points and class labels. Using Bayes’ rule, the DGM also works as a supervised classifier [24]–[25]. Hence, the DGM has the aspects of both a supervised classifier and an unsupervised feature-extractor, and thereby, it has the potential to overcome the difficulties that both conventional supervised classifiers and unsupervised feature-extractors encounter.

Given the above, this paper proposes a machine-learning-based method of diagnosing psychiatric disorders using a DGM of rs-fMRI images. Our proposed DGM considers three factors: an fMRI image, a class label (controls or patients), and a scan-wise nuisance component (signal of no interest, e.g., what a subject has in mind at that moment). Each subject is expected to belong to one of the classes. Each scan image
is independently drawn from a prior distribution \( p(z) \). Given the above, we build a scan-wise conditional generative model \( p_\theta \) of fMRI images \( x_{i,t} \) parameterized by \( \theta \). This is depicted in Fig. 1 and expressed as

\[
p_\theta(x_{i,t}|y_i) = \sum_{z_{i,t}} p_\theta(x_{i,t}, z_{i,t}|y_i) = \sum_{z_{i,t}} p_\theta(x_{i,t}|z_{i,t}, y_i)p(z_{i,t}).
\]

Based on the variational method [27], the model evidence \( \log p_\theta(x_{i,t}|y_i) \) is bounded using an inference model \( q_\phi(z_{i,t}|x_{i,t}, y_i) \) parameterized by \( \phi \) as

\[
\log p_\theta(x_{i,t}|y_i) = E_{q_\phi(z_{i,t}|x_{i,t}, y_i)} \left[ \log \frac{p_\theta(x_{i,t}, z_{i,t}|y_i)}{p_\theta(z_{i,t}|x_{i,t}, y_i)} \right] \\
\geq E_{q_\phi(z_{i,t}|x_{i,t}, y_i)} \left[ \log \frac{p_\theta(x_{i,t}, z_{i,t}|y_i)}{q_\phi(z_{i,t}|x_{i,t}, y_i)} \right] \\
+ D_{KL}(q_\phi(z_{i,t}|x_{i,t}, y_i)||p_\theta(z_{i,t}|x_{i,t}, y_i)) \\
\geq E_{q_\phi(z_{i,t}|x_{i,t}, y_i)} \left[ \log \frac{p_\theta(x_{i,t}, z_{i,t}|y_i)}{q_\phi(z_{i,t}|x_{i,t}, y_i)} \right] \\
- D_{KL}(q_\phi(z_{i,t}|x_{i,t}, y_i)||p(z_{i,t})) \\
+ E_{q_\phi(z_{i,t}|x_{i,t}, y_i)} \left[ \log p_\theta(x_{i,t}|z_{i,t}, y_i) \right] \\
= : \mathcal{L}(x_{i,t}; y_i),
\]

where \( D_{KL}(\cdot||\cdot) \) is the Kullback-Leibler divergence and \( \mathcal{L}(x_{i,t}; y_i) \) is the evidence lower bound. Because the fMRI images \( x_{i,t} \) are assumed to be obtained independently from each other, the subject-wise conditional generative model \( p_\theta(x_{i,t}|y_i) \) and its evidence lower bound \( \mathcal{L}(x_{i,t}; y_i) \) are simply the scan-wise sum:

\[
\log p_\theta(x_{i,t}|y_i) = \sum_{t=1}^{T_i} \log p_\theta(x_{i,t}|y_i) \\
\geq \sum_{t=1}^{T_i} \mathcal{L}(x_{i,t}; y_i) \\
= \mathcal{L}(x_{i,t}; y_i).
\]

In addition, the conditional generative model \( p_\theta(X|y) \) of the complete dataset and its evidence lower bound \( \mathcal{L}(X; y) \) are also expressed as the sum of the subject-wise models:

\[
\log p_\theta(X|y) = \sum_{i=1}^{N} \log p_\theta(x_{i,t}|y_i) \\
\geq \sum_{i=1}^{N} \mathcal{L}(x_{i,t}; y_i) \\
= \mathcal{L}(X; y).
\]

In general, the evidence lower bound \( \mathcal{L}(X; y) \) of the complete dataset is the objective function of the parameters \( \theta \) and \( \phi \) of the conditional generative model \( p_\theta \) and the inference model \( q_\phi \) to be maximized. In practice, we train the scan-wise model \( p_\theta(x_{i,t}|y_i) \) to maximize its evidence lower bound \( \mathcal{L}(x_{i,t}; y_i) \), and thereby train the conditional generative model \( p_\theta(X|y) \) of the complete dataset.

B. Diagnosis based on Generative Model

Once the conditional generative model \( p_\theta \) is trained, we can assume the class \( y \) of a test subject \( j \), who has not yet received a diagnosis. This diagnosis is based on Bayes’ rule and the evidence lower bound \( \mathcal{L}(x_{j}; y) \), which approximates
C. Deep Neural Generative Model of FMRI Images

In this section, we implement the conditional generative model \( p_\theta \) described in the previous section using deep neural networks, thus obtaining a deep neural generative model (DGM) [19]–[22]. We build and train the scan-wise model \( p_\theta(x_{i,t}|y_i) \), and thereby obtain the subject-wise model \( p_\theta(x_i|y_i) \) and the model \( p_\theta(X|y) \) of the complete dataset.

The inference model \( q_\phi(z_{i,t}|x_{i,t},y_i) \) is implemented on a neural network called encoder, depicted in the left part of Fig. 2. The encoder is given a preprocessed fMRI image \( x_{i,t} \) and the corresponding class label \( y_i \), then infers the posterior distribution \( q_\phi(z_{i,t}|x_{i,t},y_i) \) of the latent variable \( z_{i,t} \). Since the posterior distribution \( q_\phi(z_{i,t}|x_{i,t},y_i) \) is modeled as a multivariate Gaussian distribution with a diagonal covariance matrix, the encoder outputs a mean vector \( \mu_{z_{i,t}} \) and a variance vector \( \sigma_{z_{i,t}}^2 \). The conditional generative model \( \log p_\theta(x_{i,t}|y_i) \) is implemented on a neural network called decoder (or sometimes called generator), also depicted in the right part of Fig. 2. The decoder is given a class label \( y_i \) and a latent variable \( z_{i,t} \), then generates the posterior distribution \( p_\theta(x_{i,t}|z_{i,t},y_i) \) of an fMRI image \( x_{i,t} \).

More specifically, we constructed the encoder and decoder as follows. We assumed a preprocessed fMRI image \( x_{i,t} \) as an \( n_z \)-dimensional vector, a latent variable \( z_{i,t} \) as an \( n_z \)-dimensional vector, and a class label \( y_i \) as a one-hot vector. The encoder and decoder have \( n_h \) hidden layers. The \( l \)-th hidden layer consists of \( n_h \) units followed by the layer normalization [28] and the ReLU activation function [29]. Each weight parameter was initialized to a sample drawn from a Gaussian distribution \( \mathcal{N}(0, 0.02^2) \) and each bias parameter was initialized to 0. The encoder accepts an fMRI image \( x_{i,t} \) with the dropout of 0.2 of ratio \( p \) at its first hidden layer and a class label \( y_i \) at its last hidden layer. The output layer of the encoder consists of \( 2 \times n_z \) units; half of the units are followed by no activation function and used as a mean vector \( \mu_{z_{i,t}} \), and the other half of the units are followed by the exponential function and used as a variance vector \( \sigma_{z_{i,t}}^2 \). Next, the decoder accepts a sample \( z_{i,t} \) from the posterior distribution \( q_\phi(z_{i,t}|x_{i,t},y_i) \) and the class label \( y_i \) at its first hidden layer. The output layer of the decoder consists of \( 2 \times n_z \) units that are used as parameters of an \( n_z \)-dimensional Gaussian distribution that represents the posterior distribution \( p_\theta(x_{i,t}|z_{i,t},y_i) \) of the fMRI image \( x_{i,t} \) in the same way as the encoder.

The encoder and decoder were jointly trained using the Adam optimization algorithm [31] with parameters \( \alpha \in \{10^{-3}, 10^{-4}\} \), \( \beta_1 = 0.9 \), and \( \beta_2 = 0.999 \). We selected hyper-parameters from \( p \in \{0, 0.5\} \), \( n_h \in \{50, 100, 200, 400\} \), and \( n_z \in \{5, 10, 20, 50, 100\} \) for \( n_h > n_z \). Note that, while deeper and deeper convolutional and recurrent neural networks are attracting increasing attention (e.g., [32], [33]), recent state-of-the-art feedforward fully-connected neural networks have one or two hidden layers [19]–[22] and a deeper network architecture is not always helpful [34]. Hence, we set the number \( n_h \) of hidden layers to two. We adjusted the imbalance in the classes via oversampling; hence, we assumed the prior probabilities \( p(y) \) of classes \( y \) as \( p(y=0) = p(y=1) = 0.5 \). We stopped the learning procedure early if the training accuracy reached 100%.

D. Data Acquisition and Preprocessing

In this study, we used a dataset of rs-fMRI images obtained from patients with schizophrenia or bipolar disorder. These data were obtained from the OpenfMRI database. Its accession number is ds000030 (https://openfmri.org/dataset/ds000030). We used all the available subjects in the dataset: 50 patients with schizophrenia, 49 patients with bipolar disorder, and 122 normal control subjects. The environmental settings were repetition time (TR) = 3000 ms, acquisition matrix size = 64 × 64 × 34, 152 images, and voxel thickness = 3.0 mm.

We performed a preprocessing procedure for rs-fMRI using the SPM12 software package obtained from http://www.fil.ion.ucl.ac.uk.
III. EXPERIMENTS AND RESULTS

A. Comparative Approaches

For comparison, we evaluated a multilayer perceptron (MLP), logistic regression (LR), and support vector machine (SVM) with or without unsupervised dimension-reduction, and Gaussian mixture model (GMM).

The MLP accepted a single image $x_{i,t}$ at once. It had $u_h$ hidden layers, each of which consisted of $n_h$ units followed by the layer normalization [28] and the ReLU activation function [29], the same as our proposed DGM. It also had an output unit followed by the logistic function, representing the posterior probability $q_\theta(y = 1|x_{i,t})$. The objective function to be minimized was cross-entropy $L_{c.e.} = -\sum_y I(y = y_i) \log q_\theta(y|x_{i,t})$, where $I(cond.)$ is the indicator function that returns 1 if cond. is true and 0 otherwise. The other conditions were the same as those for our proposed DGM. Once the MLP was trained, it sequentially accepted a set $x_j = \{x_{i,t}\}$ of fMRI images obtained from a subject $j$ and diagnosed the subject using the ensemble of the diagnoses for the $T_i$ images, also consistent with our proposed DGM. Then, $\sum_{t=1}^{T_i} \log q_\theta(y = 1|x_{j,t}) > \sum_{t=1}^{T_i} \log q_\theta(y = 0|x_{j,t})$ was considered to suggest that subject $j$ had the disorder $y = 1$ and vice versa.

When the number of hidden layers is zero, the MLP is simply called a perceptron or logistic regression (LR). We trained the LR using Newton’s method instead of gradient ascent algorithms.

The SVM accepted a single image $x_{i,t}$ and outputted a binary value representing the estimated class using linear kernels. The estimated class for a subject $i$ was determined by majority voting of $T_i$ estimations, consistent with our proposed DGM and the MLP. We selected the hyper-parameter $C$ as a compromise between training classification accuracy and margin maximization from $C \in \{\ldots, 0.1, 0.2, 0.5, 1, 2, 5, 10, \ldots\}$.

For unsupervised feature-extractors, we employed a principal component analysis (PCA) and an autoencoder (AE). We selected the number $d$ of principle components of the PCA from $d \in \{5, 10, 20, 50, 100\}$. We selected hyper-parameters of AE from $p \in \{0.0, 0.5\}$, $u_h = 2$, $n_h \in \{50, 100, 200, 400\}$, and $n_z \in \{5, 10, 20, 50, 100\}$ for $n_h > n_z$, consistent with the proposed DGM. We used mean-squared-errors for evaluating reconstruction by AE. The other conditions for AE were the same as those in our proposed DGM and the MLP. For the combination of the AE and MLP, we used the same number $n_h$ of hidden units for both the AE and MLP in order to suppress its relatively high dimensional hyperparameter-space.

The generative model of fMRI images described in Section II-A can be implemented using other generative models. For comparison, we evaluated a GMM with a diagonal covariance matrix. This GMM can be considered as a single-layer version of the proposed DGM but it is trained using Expectation-Maximization (EM) algorithm. We trained two GMMs $p_\theta(x_{i,t}, y = 1)$ and $p_\theta(x_{i,t}, y = 0)$: one for patients with disorder ($y = 1$) and the other for normal control subjects ($y = 0$). Then, we diagnosed the subject as described in Section II-B also consistent with our proposed DGM. We selected the number $n$ of mixture components of the GMM from $n \in \{1, 2, 5, 10, 20, 50, 100\}$.

B. Results of Diagnosis

Let TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively. Then, we used several measures, namely accuracy (ACC), sensitivity (SEN), specificity (SPEC), positive predictive value (PPV), and negative predictive value (NPV), defined as

\[
\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},
\]

\[
\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}},
\]

\[
\text{SPEC} = \frac{\text{TN}}{\text{TN} + \text{FP}},
\]

\[
\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}},
\]

\[
\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}.
\]

However, large values of these measures do not always indicate a good model. Recall that the dataset contains many more controls than patients. If the model estimated all the subjects were controls, the diagnosis accuracy would reach 70%. Hence, we primarily used the following balanced measures: balanced accuracy (BACC), F1 score (F1), and Matthews correlation coefficient (MCC), defined as

\[
\text{BACC} = \frac{\text{SEN} + \text{SPEC}}{2},
\]

\[
\text{F1} = \frac{2}{\text{SEN}^{-1} + \text{PPV}^{-1}},
\]

\[
\text{MCC} = \frac{(\text{TP} \times \text{TN} - \text{FP} \times \text{FN})}{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}.\]

We performed 10 trials of 10-fold cross-validations and selected hyper-parameters with which the models achieved the best balanced accuracy (BACC), as summarized in Table I for schizophrenia and in Table II for bipolar disorder. The proposed DGM achieved the best results for all the balanced measures by obvious margins and a 10-model ensemble of the proposed DGM achieved even better results. The second-best results are also emphasized in bold.
TABLE I
SELECTED HYPER-PARAMETERS AND DIAGNOSIS ACCURACIES FOR SCHIZOPHRENIA.

| Model                | Selected Hyper-Parameters | | Balanced Measures | | Other Measures |
|----------------------|---------------------------|-------------------------------|------------------|------------------|
|                      |                           | Feature Extractor | Classifier | BACC | MCC | F1 | ACC | SPEC | SEN | PPV | NPV |
| chance level         | —                         | —                | —          | —    | —   | —  | —   | —    | —   | —   | —   |
| SVM                  | —                         | —                | C = 0.5    | 0.500 | 0.709 |
| LR                   | —                         | —                | (no parameter) | 0.582 | 0.153 | 0.432 | 0.614 | 0.661 | 0.504 | 0.380 | 0.763 |
| MLP                  | —                         | n_h = 50, p = 0.5 | —          | 0.574 | 0.136 | 0.431 | 0.594 | 0.622 | 0.526 | 0.365 | 0.760 |
| PCA+SVM              | d = 1                     | —                | C = 0.5    | 0.570 | 0.127 | 0.440 | 0.562 | 0.551 | 0.588 | 0.351 | 0.764 |
| PCA+LR               | d = 50                    | —                | (no parameter) | 0.569 | 0.126 | 0.430 | 0.580 | 0.598 | 0.540 | 0.357 | 0.759 |
| PCA+MLP              | d = 50                    | n_h = 50, p = 0.5 | —          | 0.747 | 0.490 | 0.636 | 0.751 | 0.756 | 0.738 | 0.606 | 0.886 |
| AE+SVM               | n_h = 400, n_z = 50       | C = 1.0, p = 0.5 | —          | 0.699 | 0.364 | 0.576 | 0.691 | 0.679 | 0.718 | 0.481 | 0.853 |
| AE+LR                | n_h = 400, n_z = 50       | p = 0.5          | —          | 0.675 | 0.322 | 0.549 | 0.673 | 0.669 | 0.682 | 0.460 | 0.834 |
| AE+MLP               | n_h = 50, n_z = 50, p = 0.5 | n_h = 50, p = 0.5 | —          | 0.738 | 0.497 | 0.617 | 0.779 | 0.837 | 0.640 | 0.667 | 0.858 |
| GMM                  | n = 20                    | —                | —          | 0.752 | 0.509 | 0.650 | 0.799 | 0.864 | 0.640 | 0.661 | 0.803 |
| DGM (proposed)       | —                         | n_h = 200, n_z = 20, p = 0.5 | —          | 0.766 | 0.551 | 0.660 | 0.809 | 0.869 | 0.664 | 0.711 | 0.870 |
| DGM (proposed, ensemble) | —                   | n_h = 200, n_z = 20, p = 0.5 | —          | 0.777 | 0.576 | 0.680 | 0.825 | 0.894 | 0.660 | 0.744 | 0.868 |

TABLE II
SELECTED HYPER-PARAMETERS AND DIAGNOSIS ACCURACIES FOR BIPOLAR DISORDER.

| Model                | Selected Hyper-Parameters | | Balanced Measures | | Other Measures |
|----------------------|---------------------------|-------------------------------|------------------|------------------|
|                      |                           | Feature Extractor | Classifier | BACC | MCC | F1 | ACC | SPEC | SEN | PPV | NPV |
| chance level         | —                         | —                | —          | —    | —   | —  | —   | —    | —   | —   | —   |
| SVM                  | —                         | —                | C = 0.1    | 0.500 | 0.713 |
| LR                   | —                         | —                | (no parameter) | 0.505 | 0.015 | 0.136 | 0.681 | 0.921 | 0.088 | 0.312 | 0.714 |
| MLP                  | —                         | n_h = 400, p = 1.0 | —          | 0.505 | 0.010 | 0.357 | 0.533 | 0.554 | 0.457 | 0.293 | 0.716 |
| PCA+SVM              | d = 50                    | —                | C = 0.1    | 0.507 | 0.024 | 0.130 | 0.687 | 0.932 | 0.082 | 0.328 | 0.75 |
| PCA+LR               | d = 100                   | —                | (no parameter) | 0.545 | 0.071 | 0.410 | 0.543 | 0.540 | 0.551 | 0.326 | 0.748 |
| PCA+MLP              | d = 10                    | n_h = 200, p = 0.5 | —          | 0.575 | 0.152 | 0.413 | 0.576 | 0.576 | 0.573 | 0.370 | 0.776 |
| AE+SVM               | n_h = 400, n_z = 50       | C = 0.1, p = 0.5 | —          | 0.583 | 0.150 | 0.450 | 0.603 | 0.651 | 0.604 | 0.358 | 0.778 |
| AE+LR                | n_h = 200, n_z = 50       | p = 0.5          | —          | 0.578 | 0.142 | 0.445 | 0.569 | 0.556 | 0.600 | 0.354 | 0.774 |
| AE+MLP               | n_h = 200, n_z = 50, p = 0.5 | n_h = 200, p = 0.5 | —          | 0.582 | 0.180 | 0.399 | 0.634 | 0.704 | 0.460 | 0.432 | 0.773 |
| GMM                  | n = 2                     | —                | —          | 0.563 | 0.114 | 0.420 | 0.575 | 0.593 | 0.532 | 0.346 | 0.758 |
| DGM (proposed)       | —                         | n_h = 400, n_z = 50, p = 0.5 | —          | 0.611 | 0.220 | 0.450 | 0.652 | 0.709 | 0.512 | 0.434 | 0.784 |
| DGM (proposed, ensemble) | —                   | n_h = 400, n_z = 50, p = 0.5 | —          | 0.641 | 0.281 | 0.485 | 0.695 | 0.770 | 0.511 | 0.487 | 0.798 |

Fig. 3. Top 10 contributing regions and their contribution weights for schizophrenia, defined in Eq. 7.

Fig. 4. Top 10 contributing regions and their contribution weights for bipolar disorder, defined in Eq. 7.
C. Reconstruction of Signals and Contributing Regions

In the previous section, we diagnosed subjects successfully following Eq. (5), i.e., using the difference in the subject-wise evidence lower bound \( L(x_i; y) \) between the given class labels \( y \). In this section, we visualize the regions contributing to the diagnoses. The scan-wise evidence lower bound \( L(x_i; y) \) is equal to the log-likelihood \( \mathbb{E}_{q_\phi}(\log p_\theta(x_{i,t}|z_{i,t}, y)) \) of an fMRI image \( x_{i,t} \) minus the Kullback-Leibler divergence \( D_{KL}(q_\phi(z_{i,t}|x_{i,t}, y)||p(z_{i,t})) \). From the perspective of a neural network, the former is the negative reconstruction error of an autoencoder and the latter is a regularization term [19]–[22]. Here, we explicitly denote an fMRI image \( x_{i,t} \) as the set of the region-wise signals \( x_{i,t,k} \) for \( k \in \{1, 2, \ldots, 116\} \) and introduce a region-wise reconstruction error given a class label \( y \):

\[
\mathcal{W}(i, t, k; y) = -\mathbb{E}_{q_\phi}(\log p_\theta(x_{i,t,k}|z_{i,t}, y)) - \mathbb{E}_{q_\phi}(\log p_\theta(x_{i,t,k}|z_{i,t}, y)) - \log 2\pi
\]

When the reconstruction error of a region \( k \) becomes much larger given the incorrect class label, the proposed DGM disentangles the signals of the region \( k \) obtained from controls and patients and the region \( k \) contributes largely to the correct diagnosis. Hence, we define the contribution weight \( \mathcal{W}(k) \) of a region \( k \) as

\[
\mathcal{W}(k) = \mathbb{E}_{i,t} \left[ \mathcal{W}(i, t, k; y = 1 - y_i) - \mathcal{W}(i, t, k; y = y_i) \right].
\]

IV. DISCUSSION

A. Diagnosis Accuracy

We summarize the selected hyper-parameters and the results in Tables III and IV. With feature extraction via PCA, the SVM and LR (noted as PCA+SVM and PCA+LR) produced a bit worse \( F_1 \) scores and MCCs in schizophrenia dataset and they produced a bit better results in bipolar disorder dataset than without the PCA. The PCA extracted features using linear kernels, which implies that the PCA played a similar role to the SVM and LR and is therefore almost redundant. Worse, features could potentially be lost in nuisance components for bipolar disorder. In contrast, feature extraction via AE improved the results of both the SVM and LR (see AE+SVM and AE+LR) because, using the multilayer architecture, the AE
extracted features whose elements were almost independent from each other and were easily classified by the SVM and LR [36]. However, the AE+SVM and AE+LR were not superior to the MLP and the proposed DGM for the schizophrenia dataset. Recall that our experimental setting allowed the AE+SVM and AE+LR to have larger numbers of parameters than the MLP and proposed DGM with the selected hyperparameters (see Section III-A). That is, the AE+SVM and AE+LR have sufficient complexity. The difference is whether the methods employed an end-to-end approach or not. Unsupervised dimension-reductions extract salient components from an fMRI image, but such components are not guaranteed to be brain activity that is related to the disorder and thus do not necessarily contribute to a correct diagnosis. They could be nuisance components such as body motion and brain shape, which were not removed successfully by preprocessing. This is why the AE+SVM and AE+LR produced worse results than the MLP and proposed DGM. In contrast, MLP without feature extraction did not work well for the bipolar-disorder dataset. These results suggest that the features of schizophrenia are relatively easily captured by discriminative approaches, and the features of bipolar disorder are sensitive to generative approaches rather than discriminative approaches.

In the both datasets, the MLP without feature extraction achieved the highest specificity and the lowest sensitivity, which implies that its predictions were biased toward controls \( y = 0 \) in spite of the adjustment for the imbalance by oversampling. For classification, a discriminative model must find a certain pattern in fMRI images \( x \) that is highly related to the class label \( y \); the model sometimes finds a confounding factor, which results in overfitting. The PCA+MLP and AE+MLP achieved better balanced accuracies, MCCs, and sensitivities in addition to lower specificities, which implies the PCA and the AE prevented the MLP from producing biased predictions. Moreover, in the bipolar-disorder dataset, the PCA+MLP used a smaller number of hidden units (\( n_h = 200 \)) than the MLP without the PCA (\( n_h = 400 \)). The PCA reduced the number of parameters in the MLP and potentially prevented overfitting.

While appropriate unsupervised feature-extractions improved the accuracies of the classifiers, the proposed DGM achieved the best results for all of the balanced measures (balanced accuracy, \( F_1 \) score, and MCC) in both datasets. Unlike the AE, which carries a risk of extracting and modeling salient but nuisance components instead of a component of interest, the proposed DGM is explicitly given the class label \( y \) in addition to an fMRI image \( x \) and extracts only nuisance components \( z \). Unlike the MLP, which has a risk of finding only a confounding factor in an fMRI image \( x \), the proposed DGM must model the whole fMRI image \( x \). Hence, the proposed DGM overcomes both difficulties of the discriminative and generative models and improves diagnosis accuracy.

The same holds true for the GMM, which can be considered as a single-layer version of the proposed DGM. The GMM achieved the second best results for the schizophrenia dataset but it achieved worse results for the bipolar-disorder dataset than the PCA+MLP and the AE+SVM/LR/MLP. These results suggest that the features of schizophrenia are relatively discriminable by linear models and the features of bipolar disorder require non-linear bases.

B. Reconstruction of Signals and Contributing Regions

As summarized in Fig. 3, the proposed DGM found that the signals obtained from the thalamus (ThalamusL and ThalamusR) significantly contributed to the correct diagnoses of schizophrenia. This result agrees with many previous studies, which have demonstrated the relationship between schizophrenia and the thalamus [37], [38]. The proposed DGM also identified brain regions that have been highlighted as related to schizophrenia in previous studies: the cerebellar vermis (Vermis) [39], the parahippocampal gyrus (ParaHippocampal) [40], the superior temporal gyrus (Temporal_Sup) [41], [42], and gyrus rectus (Rectus) [43]. In the diagnosis of bipolar disorder summarized in Fig. 4, the proposed DGM also found several significant regions that have been mentioned in previous studies: e.g., cerebellum [44], [45], frontal lobe including the inferior frontal gyrus (Frontal_Inf_Orb_R) and superior temporal gyrus (Frontal_Sup_R) [46], [47], and thalamus [48]. Therefore, we conclude that the proposed DGM successfully identified the brain regions related to each disorder and can motivate further biological investigations.

Several studies have already employed DNNs for diagnosis of neurological and psychiatric disorders and have attempted to identify regions and activity related to disorders. In [7], the contributing regions were identified based on the weight parameters of the AE: If several units representing brain regions in the input layer were connected to the same unit in the first hidden layer via large weight parameters, the regions were considered to contribute to the diagnosis with large weights. However, if such a hidden unit had another largely biased input or a large bias parameter, the unit would be saturated after the activation function and could not transfer meaningful information to the subsequent layer, i.e., the unit would be “dead” (see Chapter 6, [49]). A hidden unit also does not function when it is connected to the next hidden unit via a near-zero weight parameter. Unlike PCA, the DNNs extract nonlinear and higher-order features not only in the first hidden layer but also in the subsequent layers. The units and layers where features are extracted and whether the extracted features are actually used in the following layers are essentially uncontrolled [50], [51]. As such, one cannot quantitatively compare contribution weights between multiple input units. Conversely, the proposed DGM used region-wise reconstruction errors for the diagnosis. Hence, the regions with large reconstruction errors certainly contributed to the diagnosis and the reconstruction errors corresponded to the contribution weights of the regions.

Note that the proposed DGM is not robust to correlated regions since it assumes a diagonal covariance matrix for the posterior distribution \( p_\theta(x_{i|t}, z_{i|t}, y) \). For example, when the left thalamus is parcellated into two regions, each of them has a similar influence on the diagnosis. DGMs with non-zero covariances and generative models without explicit
distributions (e.g., generative adversarial networks [52, 53]) could overcome this issue. DGMs with temporal dynamics could also relax the assumption that the nuisance components \( z \) are independent of each other.

As shown in Fig. 5, the reconstructed time-series were apparently similar to the originals, regardless of the class label. The reconstruction error \( \mathcal{W}(i, t, k; y) \) of the right pre-cuneus (which has the lowest contribution weight) was almost independent from the given class label. In contrast, the reconstruction error \( \mathcal{W}(i, t, k; y) \) of the left thalamus was certainly smaller with the correct label \( y \) = 1 than with the incorrect label \( y \) = 0. The proposed DGM found a small but clear difference in the fMRI images between patients and controls despite the training procedure, in which the proposed DGM was not trained to discriminate these two entities. The proposed DGM can be trained as a discriminative model: Such a learning procedure increases the risk of overfitting but potentially contributes to further improvements [23, 25].

V. CONCLUSION

This study proposed a deep neural generative model (DGM) for diagnosing psychiatric disorders. The DGM was a generative model implemented using deep neural networks. The DGM modeled the joint distribution of rs-fMRI images, class labels, and nuisance components. Using Bayes’ rule, the DGM diagnosed test subjects with higher accuracy than other competitive approaches: logistic regression, support vector machine, and multilayer perceptron with or without unsupervised feature-extractors. In addition, the DGM visualized brain regions that contribute to accurate diagnoses, which motivates further biological investigations.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Ben Seymour, Dr. Kenji Leibnitiz, Dr. Hiroaki Mano, and Dr. Ferdinand Peper at CiNet for valuable discussions. This study was partially supported by the JSPS KAKENHI (16K12487), Kayamori Foundation of Information Science Advancement, and SEI Group CSR Foundation.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

[1] T. J. Sejnowski, P. S. Churchland, and J. A. Movshon, “Putting big data to good use in neuroscience,” Nature Neuroscience, vol. 17, no. 11, pp. 1440–1441, 2014.

[2] B. D. W. Group, “Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework,” Clinical Pharmacology and Therapeutics, vol. 69, no. 3, pp. 89–95, 2001.

[3] S. M. Plos et al., “Deep learning for neuroimaging: a validation study,” Frontiers in Neuroscience, vol. 8, no. Aug., pp. 1–11, 2014.

[4] E. Castro et al., “Deep Independence Network Analysis of Structural Brain Imaging: Application to Schizophrenia,” IEEE Transactions on Medical Imaging, vol. 35, no. 7, pp. 1729–1740, 2016.

[5] S. Liu et al., “Multimodal Neuroimaging Feature Learning for Multiclass Diagnosis of Alzheimer’s Disease,” IEEE Transactions on Biomedical Engineering, vol. 62, no. 4, pp. 1132–1140, 2015.

[6] F. Li et al., “A Robust Deep Model for Improved Classification of AD/MCI Patients,” IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 5, pp. 1610–1616, 2015.

[7] H.-I. Suk et al., “State-space model with deep learning for functional dynamics estimation in resting-state fMRI,” NeuroImage, vol. 129, pp. 292–307, 2016.

[8] C. Salvatore et al., “Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer’s disease: a machine learning approach,” Frontiers in Neuroscience, vol. 9, no. Sep., pp. 1–13, 2015.

[9] H.-I. Suk, S-W. Lee, and D. Shen, “Deep ensemble learning of sparse regression models for brain disease diagnosis,” Medical Image Analysis, vol. 37, pp. 101–113, 2017.

[10] L. Zhang et al., “Machine learning for clinical diagnosis from functional magnetic resonance imaging,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 1211–1217 vol. 1, 2005.

[11] S. Vergun et al., “Classification and extraction of resting state networks using healthy and epilepsy fMRI data,” Frontiers in Neuroscience, vol. 10, no. Sep., pp. 1–13, 2016.

[12] X. Guo et al., “Dиagnosing Autism Spectrum Disorder from Brain Resting-State Functional Connectivity Patterns Using a Deep Neural Network with a Novel Feature Selection Method,” Frontiers in Neuroscience, vol. 11, no. Aug., pp. 1–19, 2017.

[13] A. Abraham et al., “Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example,” NeuroImage, vol. 147, no. 15 February 2017, pp. 736–745, 2017.

[14] Y. Shimizu et al., “Toward probabilistic diagnosis and understanding of depression based on functional MRI data analysis with logistic group LASSO,” PLOS ONE, vol. 10, no. 5, pp. 1–23, 2015.

[15] B. Biswal et al., “Functional connectivity in the motor cortex of resting human brain using echo-planar MRI,” Magnetic Resonance in Medicine, vol. 34, no. 4, pp. 537–41, 1995.

[16] Z. Lin et al., “Simultaneous dimension reduction and adjustment for confounding variation,” Proceedings of the National Academy of Sciences, vol. 113, no. 51, pp. 14662–14667, 2016.

[17] Y. Bengio, A. Courville, and P. Vincent, “Representation Learning: A Review and New Perspectives,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

[18] J. Schmidhuber, “Deep learning in neural networks: An overview,” Neural Networks, vol. 61, pp. 85–117, 2015.

[19] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” International Conference on Learning Representations (ICLR), pp. 1–14, 2014.

[20] D. P. Kingma, D. J. Rezende, and M. Welling, “Semi-supervised Learning with Deep Generative Models,” in Advances In Neural Information Processing Systems (NIPS), 2014, pp. 3581–3589.

[21] K. Sohn, H. Lee, and X. Yan, “Learning Structured Output Representation using Deep Conditional Generative Models,” in Advances In Neural Information Processing Systems (NIPS), 2015, pp. 3483–3491.

[22] L. Maaløe et al., “Auxiliary Deep Generative Models,” in International Conference on Machine Learning (ICML), vol. 48, 2015, pp. 1–5.

[23] J. Lasserre, C. Bishop, and T. Minka, “Principled Hybrids of Generative and Discriminative Models,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, no. 6. IEEE, 2006, pp. 87–94.

[24] A. Prasad, A. Niculescu-Mizil, and P. K. Ravikumar, “On Separability of Loss Functions, and Revisiting Discriminative Vs Generative Models,” in Advances in Neural Information Processing Systems, 2017, pp. 7055–7062.

[25] T. Mutasbara, R. Akita, and K. Uehara, “Stock Price Prediction by Deep Neural Generative Model of News Articles,” IIEICE Transactions on Information and Systems, p. accepted, 2018.

[26] T. Tashiro, T. Mutasbara, and K. Uehara, “Deep Neural Generative Model for fMRI Image Based Diagnosis of Mental Disorder,” in International Symposium on Nonlinear Theory and its Applications (NOLTA), 2017, p. accepted.

[27] K. Murphy, Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.

[28] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer Normalization,” arXiv, pp. 1–14, 2016.

[29] V. Nair and G. E. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines,” in International Conference on Machine Learning (ICML), 2010, pp. 807–814.

[30] N. Srivastava et al., “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” Journal of Machine Learning Research, vol. 15, pp. 1929–1958, 2014.

[31] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” International Conference on Learning Representations (ICLR), pp. 1–15, 2015.
32] K. He et al., “Deep Residual Learning for Image Recognition,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
33] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” pp. 1-15, 2014.
34] A. Nøkland, “Direct Feedback Alignment Provides Learning in Deep Neural Networks,” in Advances In Neural Information Processing Systems (NIPS), vol. 23, no. 3, 2016, pp. 540–552.
35] N. Tzourio-Mazoyer et al., “Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain.” NeuroImage, vol. 15, no. 1, pp. 273–289, 2002.
36] Q. V. Le et al., “Building high-level features using large scale unsupervised learning,” in International Conference on Machine Learning (ICML), 2012, pp. 507–514.
37] A. M. Brickman et al., “Thalamus size and outcome in schizophrenia.” Schizophrenia Research, vol. 71, no. 2-3, pp. 473–484, 2004.
38] D. C. Glahn et al., “Meta-Analysis of Gray Matter Anomalies in Schizophrenia: Application of Anatomic Likelihood Estimation and Network Analysis,” Biological Psychiatry, vol. 64, no. 9, pp. 774–781, 2008.
39] T. Ichimiya et al., “Reduced volume of the cerebellar vermis in neuroleptic-naive schizophrenia.” Biological Psychiatry, vol. 49, no. 1, pp. 20–27, 2001.
40] K. Sim et al., “Hippocampal and parahippocampal volumes in schizophrenia: A structural MRI study,” Schizophrenia Bulletin, vol. 32, no. 2, pp. 332–340, 2006.
41] L. Xu et al., “Source-based morphometry: The use of independent component analysis to identify gray matter differences with application to schizophrenia.” Human Brain Mapping, vol. 30, no. 3, pp. 711–724, 2009.
42] C. N. Gupta et al., “Patterns of gray matter abnormalities in schizophrenia based on an international mega-analysis,” Schizophrenia Bulletin, vol. 41, no. 5, pp. 1133–1142, 2015.
43] G.-W. Kim, Y.-H. Kim, and G.-W. Jeong, “Whole brain volume changes and its correlation with clinical symptom severity in patients with schizophrenia: A DARTEL-based VBM study,” Plos One, vol. 12, no. 5, p. e0177251, 2017.
44] C. P. Johnson et al., “Brain abnormalities in bipolar disorder detected by quantitative T1ρ mapping,” Molecular Psychiatry, vol. 20, no. 2, pp. 201–206, 2015.
45] C. Laidi et al., “Cerebellar volume in schizophrenia and bipolar I disorder with and without psychotic features.” Acta Psychiatrica Scandinaiva, vol. 131, no. 3, pp. 223–233, 2015.
46] M. Kameyama et al., “Frontal lobe function in bipolar disorder: A multichannel near-infrared spectroscopy study.” NeuroImage, vol. 29, no. 1, pp. 172–184, 2006.
47] G. Roberts et al., “Functional Dysconnection of the Inferior Frontal Gyrus in Young People With Bipolar Disorder or at Genetic High Risk,” Biological Psychiatry, vol. 81, no. 8, pp. 718–727, 2017.
48] K. Radenbach et al., “Thalamic volumes in patients with bipolar disorder,” European Archives of Psychiatry and Clinical Neuroscience, vol. 260, no. 8, pp. 601–607, 2010.
49] A. Gibson and J. Patterson, Deep Learning A Practitioner’s Approach. O’Reilly Media, 2017.
50] S. Zhao, J. Song, and S. Ermon, “Learning Hierarchical Features from Generative Models,” in International Conference on Machine Learning (ICML), vol. 70, feb 2017, pp. 4091–4099.
51] A. Ghorbani, A. Abid, and J. Zou, “Interpretation of Neural Networks is Fragile,” arXiv, pp. 1–18, 2017.
52] I. J. Goodfellow et al., “Generative Adversarial Nets,” in Advances In Neural Information Processing Systems (NIPS), 2014, pp. 2672–2680.
53] D. Tran, R. Ranganath, and D. M. Blei, “Hierarchical Implicit Models and Likelihood-Free Variational Inference,” in Advances in Neural Information Processing Systems, feb 2017, pp. 5529–5539.