# Deep Causality Variational Autoencoder Network for Identifying the Potential Biomarkers of Brain Disorders

Amani Alfakih, Zhengwang Xia, Bahzar Ali, Saqib Mamoon, and Jianfeng Lu, *Member, IEEE*

***Abstract*—Identifying causality from observational time-series data is a key problem in dealing with complex dynamic systems. Inferring the direction of connection between brain regions (i.e., causality) has become the central topic in the domain of fMRI. The purpose of this study is to obtain causal graphs that characterize the causal relationship between brain regions based on time series data. To address this issue, we designed a novel model named deep causal variational autoencoder (CVAE) to estimate the causal relationship between brain regions. This network contains a causal layer that can estimate the causal relationship between different brain regions directly. Compared with previous approaches, our method relaxes many constraints on the structure of underlying causal graph. Our proposed method achieves excellent performance on both the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Autism Brain Imaging Data Exchange 1 (ABIDE1) databases. Moreover, the experimental results show that deep CVAE has promising applications in the field of brain disease identification.**

***Index Terms*—Causal inference, fMRI, autoencoder, Alzheimer's disease (AD), autism spectrum disorder (ASD).**

## I. Introduction

FUNCTIONAL brain networks, which describe the intricate patterns between different brain regions, have been extensively utilized in the detection of neurological disorders. Many neurological diseases have been shown to be closely related to abnormal functional connectivity between brain regions, such as Alzheimer's disease (AD) [1], [2], Parkinson's disease (PD) [3], and autism spectrum disorder (ASD) [4].

Functional Connectivity (FC) is usually defined as the degree of temporal correlation between brain regions, and some studies have shown that FC contains rich dynamic temporal information. Leonardi et al. [5] found that non-stationary functional connectivity could reflect rich and additional information about the organization of the brain. Damaraju et al. [6] used the entire time series and sliding time windows to identify schizophrenia disease, and advocated dynamic analysis to better understand the pathogenesis of brain diseases. Therefore, the dynamic functional network study can help to further discover the working mode of the brain, and can understand the brain's functional organization, which will be very useful in diagnosing brain diseases.

Recent decades, researchers have focused on constructing brain functional networks from purely observational data [7], [8]. For example, Liao et al. [9] proposed a novel brain network construction method and used it for the identification of anxiety disorder, and found that the functional connectivity between many brain regions was abnormal in patients. To distinguish Alzheimer's disease patients from healthy controls, Hojjati et al. [10] combined machine learning and graph theory to identify changes in functional brain networks in patients with mild cognitive impairment. Qiao et al. [11] introduced modular prior knowledge in the process of building a brain network, and further transformed it into a sparse low-rank graph learning problem, which can be solved by machine learning algorithms. Wang et al. [12] systematically investigated the key techniques required for the diagnosis of brain diseases, including the construction of functional brain networks, brain network analysis, and a wide variety of classification methods.

Recently, causal discovery has flourished in many branches of science, such as formulating and testing hypotheses, interpreting data, prioritizing experiments, and improving or building models or theories [13]. This is also true in the domain of fMRI research, where researchers are very interested in identifying causal relationships between brain regions, i.e., the flow of signals. Researchers have developed a number of deep learning models for the identification of causality. For example, Nauta et al. [14] combined the attention mechanism with the causal validation theory in Temporal Causal Discovery

Manuscript received 26 May 2023; revised 10 September 2023 and 16 November 2023; accepted 16 December 2023. Date of publication 19 December 2023; date of current version 16 January 2024. This work was supported by the Jiangsu Natural Science Foundation under Grant BK20221487. *(Corresponding author: Jianfeng Lu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Board of ADNI and ABIDE (see http://adni.loni.usc.edu/study-design/ongoing-investigations/ and http://fcon_1000.projects.nitrc.org/indi/abide/).

Amani Alfakih is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China, and also with the Computer Science Department, Faculty of Sciences, Ibb University, Ibb, Yemen (e-mail: am775901039@gmail.com).

Zhengwang Xia, Saqib Mamoon, and Jianfeng Lu are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: xzhengwang@njust.edu.cn; saqibmamoon@njust.edu.cn; lujf@njust.edu.cn).

Bahzar Ali is with the School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: AliSalah@njust.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3344995

Framework (TCDF) to construct a new causal inference model. The model can not only identify causal relationships between variables, but also determine the time delay between the cause and its effect. Duan et al. [15] proposed a novel deep learning model to infer causal relationships between multiple variables, which were combined with graph neural networks to achieve more accurate predictions. Wein et al. [16] proposed a graph neural network model to infer causal dependencies between brain regions, and experimental results demonstrated the superiority of the method.

Commonly, the categorization of brain diseases using causal networks involves a three-step process: (1) constructing brain functional networks; (2) extracting features based on the constructed functional networks to train the classification model; (3) giving prediction results for those unknown samples based on the trained model. Several popular methods have been adopted to estimate the direction of information between brain regions. For example, Bayesian Network (BN) is a typical method for identifying causal relationships between variables. However, the method has constrained that the causal graph must be a directed acyclic graph (DAG), which limits its application [17], [18]. In addition, Yu et al. [19] designed a deep generative model for learning causal relationships among variables, and they also constrained the underlying graph structure to be a DAG. Therefore, this article proposed a unified framework that can infer the structure of brain networks in an end-to-end manner. The main contributions of this article can be summarized as follows:

- We develop a deep generative network which is a cyclic graph aiming to capture the sample distribution more accurately.
- We construct a new deep learning model called deep Causal Variation Autoencoder (deep CVAE). This network contains a causal layer that can estimate the causal relationship between different brain regions directly.
- In order to prove the effectiveness of our work, we evaluate the deep CVAE and several other competing methods on two public databases: Alzheimer's Disease Neuroimaging Initiative (ADNI), and Autism Brain Imaging Data Exchange (ABIDE). The experimental results indicate that deep CVAE performs admirably when compared with other benchmark methods.

The remainder of the article is structured as follows. In Section II, the materials and methods are presented, including the data acquisition and preprocessing, proposed framework, deep variational autoencoder, and learning strategy and loss function. Then, the experimental results and discussion are described in Sections III and IV, respectively. Finally, in Section V, the conclusion of this study is drawn.

## II. Materials and Methods

In this section, we first briefly describe the data acquisition and preprocessing process, and then provide the proposed research framework employed in this work.

### A. Data Acquisition and Preprocessing

To further validate the effectiveness of the proposed method in this paper, we conducted experiments on two publicly available datasets (Alzheimer's Disease Neuroimaging Initiative and Autism Brain Imaging Data Exchange).

TABLE I
Detailed Information of the ADNI Dataset

| Group | NC | eMCI | LMCI |
|---|---|---|---|
| Male/Female | 28/39 | 32/45 | 50/20 |
| Age(mean ±STD) | 74.1± 6.2 | 71.2±6.9 | 71.2±8.3 |

*1) ADNI Database:* The dataset contains 214 subjects with three categories, including normal controls (NC), patients with early mild cognitive impairment (eMCI), and patients with late mild cognitive impairment (LMCI). For more details regarding imaging parameters, please refer to the ADNI protocols,[1] and the detailed information is summarized in Table I.

*2) ABIDE Database:* Autism Brain Imaging Data Exchange (ABIDE) is a multi-site dataset that contains 1112 subjects from 17 different sites. All participants had corresponding functional MRI and phenotypic information. The detailed scan procedures and protocols can be referred to the ABIDE website.[2] Considering that some sites have only a small number of participants, we used data from 5 different sites in this experiment. Each site has more than 50 subjects, including Leuven, NYU, UCLA, UM, and USM. Specifically, a total of 593 subjects, including 287 patients and 306 normal controls, were enrolled. Table II lists the detailed demographic information.

In both databases, for the preprocessing of fMRI data, the following standard pipeline is adopted. In the beginning, to avoid noise signals, we discarded the first 5 volumes for each subject before preprocessing, the remaining volumes were reserved for the subsequent analysis. All functional images were transformed into the Montreal Neurological Institute (MNI) space with a resample voxel size of $3 \times 3\ \times 3$ mm$^3$. After that, Conn Toolbox 20b, a preprocessing pipeline based on Statistical Parametric Mapping, was used to perform outlier detection, direct segmentation, normalization, linear detrending, and functional smoothing with a Gaussian kernel of 8mm full width half maximum (FWHM), etc. Finally, the time series of each brain region is extracted from the preprocessed images based on the AAL atlas.

### B. Proposed Framework

The proposed framework can be divided into three steps, as illustrated in Fig. 1. First, the time series data of each brain region is extracted from the pre-processed fMRI images according to the AAL atlas. Then, we feed the time series data into a deep generative model to estimate the causal relationships between brain regions. Finally, we trained the classifier to identify patients from the enrolled subjects. The most important step is the second one, where a well-built brain network model can provide richer and more effective features for the identification of brain diseases.

[1] http://adni.loni.ucla.edu

[2] http://fcon_1000.projects.nitrc.org/indi/abide/

TABLE II
DEMOGRAPHIC INFORMATION OF SUBJECTS FROM THE FIVE SITES OF THE ABIDE DATASET

| Data Site | Total | ASD | | TC | |
|---|---|---|---|---|---|
| | | Male/Female | Age (mean ±STD) | Male/Female | Age (mean ±STD) |
| Leuven | 64 | 26/3 | 17.75 ± 4.91 | 30/5 | 18.17 ± 4.91 |
| NYU | 184 | 68/11 | 14.52 ± 6.92 | 79/26 | 15.81 ± 6.22 |
| UCLA | 99 | 47/6 | 13.01 ± 2.45 | 40/6 | 12.93 ± 1.89 |
| UM | 145 | 58/10 | 13.13 ± 2.39 | 59/18 | 14.79 ± 3.55 |
| USM | 101 | 58/0 | 22.65 ± 7.66 | 43/0 | 21.36 ± 7.55 |
| Total | 593 | 257/30 | | 251/55 | |

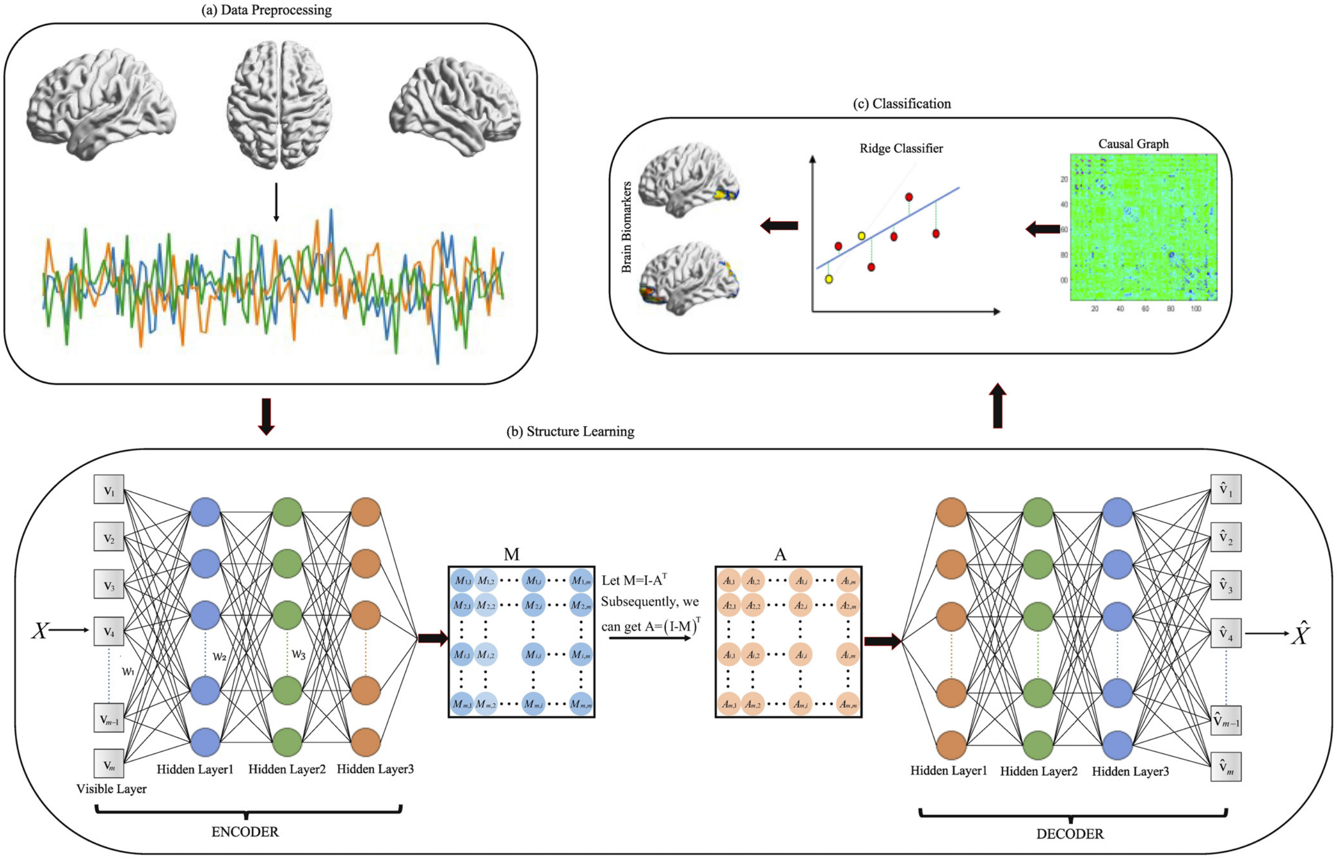


Fig. 1. The illustration for the identification of brain disorders. (a) extract time series data from fMRI images; (b) feed the time series data X into the designed network to obtain causal graph, i.e., the brain network A; (c) train classifier to identify patients with brain diseases.

## C. Deep Variational Autoencoder

Fig. 1(B) illustrates the architecture of our structure learning, which consists of two modules: an encoder and a decoder. The objective of the autoencoder is to obtain a causal representation of the input data. The fundamental principle that "causes generate results" is upheld in the field of causality [20], [21]. The concept of autoencoders is aligns with the principle of causality [22]. The encoder can be considered as tracing the underlying cause, while the decoder's objective is to reconstruct the input based on the underlying cause.

Inspired by the work of directed acyclic graph with graph neural network (DAG_GNN) [19], we have developed a new network with the specific objective of estimating causal relationships between brain regions. Specifically:

- It can run end-to-end in an unsupervised manner. This is because we cannot know whether the subject is sick at the beginning when we assist the doctor with diagnosis.
- In comparison to traditional approaches for modeling brain networks, our method excels in its ability to model nonlinearity.
- We remove the acyclic constraint from [19], which is more consistent with the physiological working mechanism of the brain [23]. Some studies have revealed the existence of loops in the human brain, which may play a role in regulating emotions and alleviating stress [24], [25].
- Additionally, we add more hidden layers to both the encoder and decoder in order to enhance representation learning and capture more accurate causal relationships.

As we described in subsection II-B, constructing a robust brain network is the most crucial step. Therefore, we have designed a new model to estimate the causal relationship between brain regions. The specific network structure is shown in Fig. 1(b). The construction of brain networks is transformed into a problem of learning graph structures.

Our approach builds upon the foundation of the linear structural equation model (SEM) and employs a deep generative model to learn the weights of a brain network. Suppose $A \in \mathbb{R}^{m\times m}$ is a weighted graph with $m$ nodes. According to the SEM theory, the causal relationship between $m$ brain regions can be expressed as:

$$X = A^T X + Y = M^{-1} Y \tag{1}$$

where $M$ is equal to $I$-$A^{\mathrm{T}}$, $I$ is the identity matrix. $Y$ is the noise matrix, which represents the independent Gaussian exogenous factors. $A$ is the brain network to be optimized. $X \in \mathbb{R}^{m\times d}$ is the time series data from the fMRI images, $d$ denotes the length of time series data. When the nodes in the graph are sorted in a topological order, matrix $A$ becomes a strictly upper triangular matrix. In this context, conducting ancestral sampling from the directed acyclic graph (DAG) can be equivalent to generating random noise $Y$ and then performing a triangular solve.

Formula (1) can be further written in the general form as follows: $X = f_A(Y)$, which can be regarded as the general expression of a graph neural network. In this expression, $Y$ represents the node features as input and $X$ represents the high level representations as output. This form can be used to write almost all graph neural networks [26], [27], [28]. Here, the structure (1) can be defined as follows:

$$X = f_1(\alpha), \quad \alpha = M^{-1}.f_2(Y) \tag{2}$$

where $f_1$ and $f_2$ are the parameterized functions for performing transforms on $X$ and $Y$, respectively. Afterward, the corresponding encoder for the generative model (2) is constructed as follows:

$$Y = f_3\left(M.f_4\left(X\right)\right) \tag{3}$$

Conceptually, the inverse roles of $f_1$ and $f_2$ in expression 2 are performed by the parameterized functions $f_3$ and $f_4$, respectively.

Once we obtained the causal representation, it goes through the Mask Layers [29] to recreate itself. It can be seen that this step is similar to the structural causal model (SCM) that shows how children are produced by the corresponding parental variables. Observe that interfering the cause will alter the effect, whereas interfering the effect does not alter the cause because the information can only flow from the last layer of the encoder into the causal layer in our network, which is consistent with the idea of causal effects.

### D. Learning Strategy and Loss Function

This section discusses the training of the Deep CVAE network. Given the distribution of $Y$ and samples $X^1,\ldots, X^n$, where $X$ refers to the sample index and $n$ represents the total number of training samples, the decoder model can be learned by maximizing the log-evidence

$$\max \log p\left(X^k\right) = \max \log\left(p\left(Y\right). \prod_{k=1}^{n} p\left(X^k\left|Y\right.\right)\right) \tag{4}$$

However, the above equation is usually unsolvable. So, variational Bayes is applied to learn a tractable distribution $q(Y|X)$ to approximate the actual posterior $p(Y|X)$. The evidence lower bound (ELBO) is as follows:

$$\begin{aligned}&\mathrm{ELBO}^k\\ &= E_{q(Y|X^k)}\left[\log p\left(X^k\left|Y\right.\right)\right] - \mathrm{KL}\left(q\left(Y\left|X^k\right.\right) \| p\left(Y\right)\right)\end{aligned} \tag{5}$$

where ELBO$^k$ deviates from the log-evidence by KL($q(Y|X^k)$|| p($Y|X^k$)) ≥0, it represents the KL-divergence between the actual posterior and the variational one.

The evidence lower bound lends itself to a VAE [30], where the encoder encodes a given sample $X^k$ into a latent variable $Y$ with density $q(Y|X^k)$. As well as the decoder tries to recreate $X^k$ from the latent variable $Y$ with density $p(X^k|Y)$. Neural networks can be used to parameterize both densities. In order to complete the deep CVAE, it is essential to provide an exhaustive description of the probability distributions outlined in equation (5). As a reminder, $X^k$ and $Y$ are both $m \times d$ matrices at now, where $m$ and $d$ represent the number of brain regions and the time series length.

For the encoder model, the identity mapping and a multilayer perceptron are employed to represent $f_3$ and $f_4$, respectively. The variational posterior $q(Y|X)$ is then represented as a Gaussian distribution, with its mean $M_Y \in \mathbb{R}^{m\times d}$ and standard deviation $S_Y \in \mathbb{R}^{m\times d}$ calculated from the inference model

$$\left[M_Y | \log S_Y\right] = M\mathrm{MLP}\left(X, W^1, W^2, W^3, W^4\right) \tag{6}$$

where MLP($X$, $W^1$, $W^2$, $W^3$, $W^4$) typically represents a Multi-Layer Perceptron (MLP) neural network with four layers (visible layer and three hidden layers). $X$ refers to the data that we feed into the neural network for processing. $W^1$, $W^2$, $W^3$ and $W^4$ represent the weight matrices connecting these layers. Moreover, the size of $W^1$ is $m \times p$, both $W^2$ and $W^3$ are $p \times p$, and $W^4$ is $p\times$13,456, $p$ is the number of hidden units.

Similarly, for the decoder model, an MLP and an identity mapping are used to represent $f_1$ and $f_2$, respectively. The likelihood $p(X|Y)$ is a factored Gaussian with mean $M_X \in \mathbb{R}^{m\times d}$ and standard deviation $S_X \in \mathbb{R}^{m\times d}$, which can be calculated from the generative model:

$$\left[M_X | \log S_X\right] = \mathrm{MLP}\left(M^{-1}Y, W^5, W^6, W^7, W^8\right) \tag{7}$$

where $W^5$, $W^6$, $W^7$ and $W^8$ represent the weight matrices connecting the decoder layers. The size of $W^5$ is 13,456$\times p$, both $W^6$ and $W^7$ are $p \times p$, and $W^8$ is $p \times m$.

In accordance with (6) and (7), the term of KL-divergence in the equation (5) can be computed based on the output of the encoder (6), which includes both $M_Y$ and $S_Y$. Moreover, the term of reconstruction accuracy $E_{q(Y|X^k)}[\log p(X^k|Y)]$ in (5) can be estimated using Monte Carlo approximation based on the output of the generative model (7).

The details of the deep CVAE network are as follows: both encoder and decoder contain 3 hidden layers, where each hidden layer consists of 100 ReLU units. For the ADNI dataset, epoch was set to 200 and batch size is set to 15. For the

ABIDE dataset, the epoch and batch size were set to 200 and 32, respectively. In the ADNI and ABIDE tasks, the model was optimized using the Adam optimizer with a learning rate of 1e-15 and 1e-5, respectively.

## III. Experiments

In this section, the compared methods are described first. Then, the specific setup of the experiment is described in detail. Finally, the classification results on the ADNI and ABIDE databases and the results of the ablation study are presented, respectively.

### A. Compared Methods

The main contribution of this paper is the construction of a deep learning model, which can estimate the causal effects between different brain regions. To further assess the effectiveness of the proposed method, we compare it with a number of popular methods. Eight popular brain network modeling methods are selected for comparison with the method proposed in this paper, including Pearson correlation-based method (PC) [4], sparse low-rank representation method (SLR) [11], Granger causality-based method (GC) [31], transfer entropy (TE) [32], linear non-Gaussian acyclic model (LiNGAM) [8] and directed acyclic graph (DAG-GNN) [19], effective temporal lag neural network (ETLN) [33] and causal recurrent variational autoencoder (CR-VAE) [34]. It is worth noting that the first two methods are based on correlation, while the last six are based on causality.

For the PC, its edge weight is defined as the temporal correlations between signals from different brain regions. For the SLR, it adds both sparsity and low-rank constraints to the brain network structure. For the GC, it determines the brain network structure by using the Granger causality test. For the TE, it calculates the causal effects between brain regions using Shannon entropy. For the LiNGAM, its core idea is that each brain region is a linear combination of all other brain regions. For the DAG-GNN, it embeds the graph structure in the graph network as a parameter to be learned for optimization. For the ETLN, the main core is the GAN structure design, which incorporates the solution target of estimation the causal relationships and temporal lag values between brain regions into the GAN model as the parameters to be learned. For the CR-VAE, the concepts of Granger causality are integrated into a recurrent VAE model.

### B. Experimental Settings and Evaluation Metrics

In order to evaluate the performance of each method more fairly, the same feature selection algorithm and classifier are adopted to test the classification performance of each brain network model. For feature selection, the weight of the network edges is seen as raw features and the recursive feature elimination (RFE) [35] method is adopted to select valid features from the raw features. For the classifier, a ridge classifier is employed for classification. For all tasks, the number of features selected by RFE is set to 4000.

For the ADNI dataset, three binary classification tasks (i.e., NC vs. eMCI, NC vs. LMCI and eMCI vs. LMCI) are performed to assess the effectiveness of the proposed approach. For the ABIDE dataset, a total of six binary classification tasks are performed. In addition, to enhance the confidence of the evaluation results, a standard 10-fold cross-validation strategy [36] was adopted. Four evaluation metrics (accuracy (ACC), sensitivity (SEN), specificity (SPE) and F1 score [37]) were adopted to evaluate the classification performance of each brain network, and the calculation of these metrics is defined as follows:

$$ACC = \frac{TN + TP}{TN + FN + TP + FP} \tag{8}$$

$$SEN = \frac{TP}{FN + TP} \tag{9}$$

$$SPE = \frac{TN}{TN + FP} \tag{10}$$

$$F1\text{score} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{11}$$

Specifically, $TP$, $TN$, $FP$ and $FN$ represent the number of true positive subjects, true negative subjects, false positive subjects and false negative subjects, respectively.

Deep neural networks can be visualized in various ways [38], [39], we choose the most commonly used layer activation method [39] for visualization because it is simple to implement. In our work, the features of the middle layer of the encoder are mapped out.

### C. Classification Results

To evaluate the performance of these methods, we perform five runs of each method and take the average of these five runs as the final result. We present the classification results of each method on the ADNI dataset and the ABIDE dataset in Table III and Table IV, respectively. The best scores are highlighted in bold. As can be observed from the two tables, our method achieves the best performance on almost all tasks, which confirm the effectiveness of our approach.

As can be seen from Table III, the proposed method achieves the best classification performance in all tasks compared to competing methods, with accuracies of 75.6%, 82.6%, and 74.4%, respectively, which demonstrates that our model is capable of extracting the causal relationships between brain regions and significantly improves the classification performance.

As can be seen from Table IV, our method performs better than competing methods on most tasks. Deep CVAE yields higher accuracy on four independent data sites, reaching 66.9%, 71.3%, 70.8%, and 76.0%, respectively. In addition, it achieves an accuracy rate of 71.4% using the whole data, which is also better than other state-of-the-art techniques. The result of NYU is somewhat inferior compared to the results of other methods, probably due to the imbalance of the data on this site.

The classification performance of LiNGAM and DAG is not satisfactory, especially on the ABIDE dataset. The reason for this may be that the brain network was added an acyclic constraint, which makes the obtained brain network too sparse and further causes the classifier to fail to capture enough effective features. The reason for the poor performance of

TABLE III
CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS ON ADNI DATABASE

| Data | Methods | ACC | SEN | SPE | F1 |
|---|---|---|---|---|---|
| NC&<br>eMCI | PC | 72.2±0.6 | **70.4±1.1** | 73.8±1.5 | 70.2±0.6 |
| | SLR | 66.1±1.7 | 61.5±2.6 | 70.1±1.6 | 62.8±2.1 |
| | GC | 65.0±3.0 | 53.7±4.5 | 74.8±1.8 | 58.8±4.1 |
| | TE | 62.5±1.0 | 58.2±2.1 | 66.2±2.3 | 59.1±1.2 |
| | LiNGAM | 63.1±1.6 | 59.1±3.8 | 66.5±3.8 | 59.8±2.1 |
| | DAG-GNN | 64.7±1.3 | 49.0±2.6 | 78.4±1.6 | 56.3±2.0 |
| | ETLN | 73.3±1.1 | 67.2±1.9 | 78.7±0.6 | 70.1±1.5 |
| | CR-VAE | 62.6±3.3 | 56.1±3.8 | 68.3±3.0 | 58.3±3.8 |
| | Deep CVAE | **75.6±1.5** | 69.2±2.1 | **81.3±2.3** | **72.4±1.6** |
| NC&<br>LMCI | PC | 78.8±0.8 | 78.8±2.4 | 78.9±2.8 | 78.4±0.8 |
| | SLR | 69.1±1.6 | 68.1±1.2 | 70.0±2.7 | 68.3±1.4 |
| | GC | 72.7±0.7 | 63.3±2.8 | 81.7±2.1 | 69.4±1.4 |
| | TE | 61.8±1.9 | 60.6±1.5 | 62.9±3.0 | 60.8±1.6 |
| | LiNGAM | 65.5±1.6 | 58.8±0.7 | 72.0±3.5 | 62.6±1.0 |
| | DAG-GNN | 57.1±1.3 | 49.9±1.5 | 64.0±1.4 | 53.2±1.4 |
| | ETLN | 78.1±0.9 | 73.4±2.4 | 82.6±1.9 | 76.6±1.2 |
| | CR-VAE | 56.4±3.2 | 50.1±2.6 | 62.3±4.1 | 52.9±3.1 |
| | Deep CVAE | **82.6±2.4** | **82.1±2.5** | **83.1±2.9** | **82.2±2.4** |
| eMCI<br>&<br>LMCI | PC | 67.5±1.3 | 67.8±2.5 | 67.1±1.3 | 68.6±1.6 |
| | SLR | 65.7±1.6 | 68.6±2.9 | 62.6±2.8 | 67.7±1.8 |
| | GC | 61.5±1.4 | 65.7±1.9 | 56.9±2.8 | 64.1±1.3 |
| | TE | 52.0±3.1 | 56.1±4.5 | 47.4±2.3 | 55.0±3.5 |
| | LiNGAM | 62.3±1.5 | 69.4±1.3 | 54.6±2.3 | 65.8±1.3 |
| | DAG-GNN | 53.5±1.2 | 63.1±4.5 | 42.9±3.6 | 58.6±2.2 |
| | ETLN | 73.5±2.5 | 77.4±3.5 | 69.1±2.9 | 75.3±2.5 |
| | CR-VAE | 58.0±1.3 | 56.0±3.8 | 55.7±2.7 | 59.9±2.2 |
| | Deep CVAE | **74.4±1.7** | **78.7±3.9** | **69.7±2.3** | **76.3±2.1** |

GC and CR-VAE may be that it constructs a binary graph from which only very few effective features can be extracted. Moreover, we can notice that the TE method obtains the worst performance in most of the tasks.

As shown in Table III and Table IV, compared with DAG-GNN, the proposed method yielded the best classification performance in most tasks. This illustrates that our model can construct better functional networks after adding the layers, which can greatly improve the classification performance. Moreover, deep CVAE has no constraints as in DAG-GNN for brain network construction. This demonstrates that our deep framework is reasonable, that is, a good classification can be obtained by employing the constructed causal graphs.

## D. Ablation Study

The main innovation of this article is to develop a model to estimate the causal effects between brain regions. To confirm whether each component helps to improve the classification performance, we designed several degraded networks for the ablation study. The three degraded networks are as follows: 1) we use only one hidden layer in each encoder and decoder model, denoted "CVAE", 2) we increase the number of hidden layers in the encoder and decoder to four, denoted "deep CVAE_1", and 3) we switch the identity mapping and MLP within each encoder/decoder, denoted "deep CVAE_2".

As can be seen from Fig. 2, the classification accuracy of the ADNI and ABIDE datasets does not improve in the increase of the number of hidden layers, which means that it is not meaningful to increase the hidden layers. Moreover, we obtained the results of the CVAE network by training it on each dataset for 100 epochs, and it achieved acceptable accuracy compared to other competing methods, especially on tasks of the ADNI database. By comparing the recognition results for the three ablation studies networks, we can find that switching identity mapping and MLP leads to degradation in the classification results, so deep CVAE_2 has the worst performance on all tasks of ADNI and ABIDE datasets. This result demonstrates that the three degraded networks are all useless to enhance the classification performance, and the deep CVAE_2 is more useless.

TABLE IV
CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS ON ABIDE DATABASE

| Data | Methods | ACC | SEN | SPE | F1 |
|---|---|---|---|---|---|
| Leuven | PC | 57.2±2.9 | 50.3±4.1 | 62.9±3.1 | 51.6±3.6 |
| | SLR | 62.8±1.2 | **56.6±1.7** | 68.0±3.3 | **58.0±0.5** |
| | GC | 59.0±3.4 | 40.7±5.4 | 75.0±3.4 | 48.0±5.1 |
| | TE | 50.0±1.0 | 41.4±3.1 | 57.1±3.1 | 42.8±1.9 |
| | LiNGAM | 59.4±1.0 | 42.1±4.0 | 73.7±2.1 | 48.3±3.0 |
| | DAG-GNN | 54.7±2.8 | 20.7±5.8 | **82.9±1.8** | 29.0±6.9 |
| | ETLN | 47.8±3.8 | 22.8±6.8 | 68.6±4.0 | 28.1±7.5 |
| | CR-VAE | 54.1±1.9 | 27.6±5.3 | 76.0±2.9 | 35.0±5.5 |
| | Deep CVAE | **66.9±2.7** | 49.7±3.5 | 81.1±2.9 | 57.6±3.5 |
| NYU | PC | **66.4±2.1** | **53.4±3.4** | 76.2±3.1 | **57.7±2.8** |
| | SLR | 63.3±0.8 | 48.4±3.1 | 74.5±1.6 | 53.0±2.0 |
| | GC | 52.0±1.2 | 21.0±3.5 | 75.2±2.4 | 27.2±3.6 |
| | TE | 58.4±2.6 | 53.2±3.3 | 62.3±3.4 | 52.3±2.9 |
| | LiNGAM | 60.0±1.6 | 45.6±1.1 | 70.9±2.5 | 49.5±1.4 |
| | DAG-GNN | 58.5±1.2 | 34.9±2.6 | 76.2±1.7 | 41.9±2.3 |
| | ETLN | 59.3±1.9 | 29.9±3.6 | **81.5±1.3** | 38.6±3.9 |
| | CR-VAE | 56.7±1.2 | 24.3±1.9 | 81.1±1.5 | 32.5±2.1 |
| | Deep CVAE | 63.9±1.0 | 43.5±2.2 | 79.2±0.7 | 50.9±1.9 |
| UCLA | PC | 66.7±1.3 | 75.1±2.2 | 57.0±2.9 | 70.7±1.2 |
| | SLR | 60.8±1.2 | 69.8±2.9 | 50.4±0.9 | 65.6±1.7 |
| | GC | 58.6±2.3 | 77.0±3.9 | 37.4±1.6 | 66.5±2.3 |
| | TE | 48.5±1.9 | 59.2±2.3 | 36.1±2.2 | 55.2±1.8 |
| | LiNGAM | 48.3±2.4 | 54.3±1.9 | 41.3±3.9 | 53.0±1.9 |
| | DAG-GNN | 52.3±2.8 | 67.2±4.2 | 35.2±4.4 | 60.1±2.6 |
| | ETLN | 48.3±2.7 | 65.3±3.5 | 28.7±3.7 | 57.5±2.5 |
| | CR-VAE | 56.8±1.6 | 68.3±1.8 | 43.5±2.7 | 62.8±1.3 |
| | Deep CVAE | **71.3±2.0** | **80.8±2.8** | **60.4±2.9** | **75.1±1.9** |
| UM | PC | 66.3±0.5 | 60.3±1.3 | 71.7±1.5 | 62.7±0.6 |
| | SLR | 67.3±1.6 | **66.5±2.2** | 68.1±2.5 | **65.6±1.6** |
| | GC | 53.0±2.3 | 37.0±3.3 | 67.2±2.5 | 42.6±3.3 |
| | TE | 58.9±0.7 | 56.8±2.0 | 60.8±2.7 | 56.4±0.8 |
| | LiNGAM | 58.8±2.2 | 54.7±3.5 | 62.3±1.8 | 55.4±2.8 |
| | DAG-GNN | 62.5±2.5 | 42.9±4.5 | 79.7±1.8 | 51.7±4.2 |
| | ETLN | 63.4±3.4 | 48.5±5.0 | 76.6±3.2 | 55.4±4.8 |
| | CR-VAE | 57.4±1.3 | 42.4±2.2 | 70.6±1.0 | 48.2±2.0 |
| | Deep CVAE | **70.8±1.8** | 58.8±3.1 | **81.3±1.0** | 65.3±2.5 |
| USM | PC | 69.3±2.1 | 72.8±2.0 | **64.7±4.0** | 73.1±1.7 |
| | SLR | 60.4±2.6 | 69.7±3.9 | 47.9±3.2 | 66.9±2.6 |
| | GC | 62.2±2.2 | **92.6±2.1** | 21.0±4.4 | 73.9±1.4 |
| | TE | 59.8±1.5 | 65.9±2.8 | 51.6±3.7 | 65.3±1.6 |
| | LiNGAM | 68.7±0.8 | 77.6±1.1 | 56.7±1.9 | 74.0±0.6 |
| | DAG-GNN | 58.6±1.9 | 89.3±2.5 | 17.2±2.4 | 71.2±1.5 |
| | ETLN | 61.0±1.5 | 85.9±1.7 | 27.4±1.7 | 71.7±1.1 |
| | CR-VAE | 53.1±1.0 | 82.4±0.7 | 13.5±2.3 | 66.9+0.6 |
| | Deep CVAE | **76.0±1.9** | 85.5±1.8 | 63.3±3.4 | **80.4±1.5** |
| Whole<br>data | PC | 66.4±0.7 | 63.3±1.2 | 69.3±1.1 | 64.6±0.8 |
| | SLR | 65.4±0.7 | 62.2±1.6 | 68.5±0.9 | 63.5±1.0 |
| | GC | 53.7±1.3 | 51.2±1.5 | 56.1±1.7 | 51.8±1.3 |
| | TE | 57.3±1.1 | 55.8±2.0 | 58.6±0.6 | 55.8±1.5 |
| | LiNGAM | 58.9±1.6 | 54.8±1.7 | 62.8±1.7 | 56.3±1.7 |
| | DAG-GNN | 58.3±1.0 | 52.2±1.8 | 63.9±1.2 | 54.7±1.4 |
| | ETLN | 61.0±0.6 | 58.3±0.9 | 63.7±0.7 | 59.1+0.7 |
| | CR-VAE | 54.4±1.2 | 47.0±1.0 | 61.3±2.2 | 50.0+1.0 |
| | Deep CVAE | **71.4±0.6** | **66.7±1.0** | **75.8±0.4** | **69.3±0.7** |

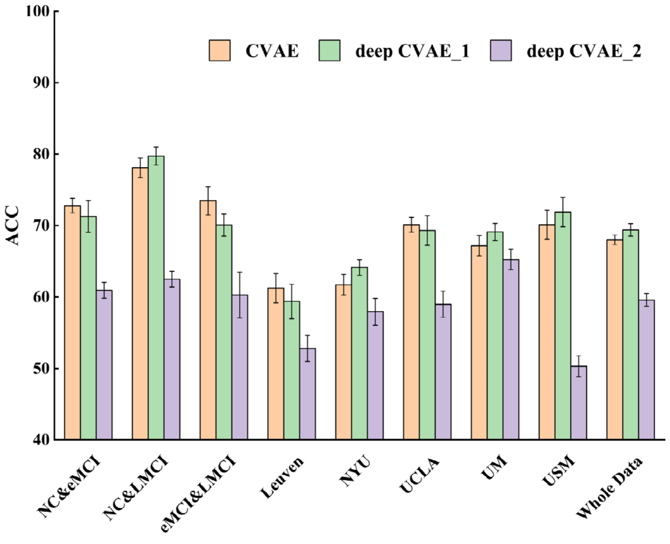


Fig. 2. The recognition results for ablation studies on ADNI and ABIDE datasets.

## IV. Discussion

In this section, we first show the most discriminative causal patterns among the six tasks. Then, we give the hierarchical causality between brain activation maps. Finally, the visualization of disease-related features is presented.

### A. Most Discriminative Patterns

Fig. 3 shows the top 30 most discriminative causal patterns across the six tasks. For the circos graph, there are three pointsto note. First, to improve the visual impact, each color of arc in the circos graph is assigned at random. Second, the significance of each connection increases as the arc width increases. Finally, the direction of ball movement in each arc represents the causal relationship between two brain regions.

By comparing the visualization results of the three tasks on the ADNI dataset, we can draw some meaningful conclusions. First, the three tasks identified multiple common brain regions as potential biomarkers of Alzheimer's disease, including CRBL3, CRBL10 and the left part of the middle temporal gyrus (TPOmid.L). Second, among of the three causal patterns, the brain region with the best discrimination was inconsistent. Third, it can be found that the classifier tends to assign higher discriminative weights to those functional connections that are farther away. We believe that this phenomenon is plausible. Multiple brain regions may be required to transmit the flow of information between distant brain regions, and abnormal changes at any node along the transmission path may result in alterations in the functional connectivity strength.

Similarly, by comparing the visualization results of the three tasks on the ABIDE database, we also obtained some interesting conclusions. First, the tasks of UCLA, UM, and USM sites jointly recognized many brain regions as potential biomarkers for ASD identification, including the right middle temporal gyrus (MTG.R), the left paracentral lobule (PCL.L) and the right inferior parietal (IPL.R). This implies that these brain regions may serve as potential biomarkers for recognizing ASD. Second, it can also be seen that the functional connections with higher weights are quite different on the three datasets. The reason may be that the ABIDE dataset contains multi-sites, and there is domain shift between the data of different sites [40].

### B. Hierarchical Causality

We investigate the relationship between brain activation maps learned in the middle layer of the inference network, and Fig. 4 shows the hierarchical relationship between brain activation maps of the different layers.

The structure of our network contains many hidden layers, each of which may capture a different causal relationship. In this section, we visualize the results of the hidden layers for all tasks. As illustrated in Fig. 4, we show the most discriminative features extracted by the middle layer of the inference model. Those areas closer to red indicate that the two groups of subjects differed more in that area. Comparing the results of ADNI, it can be discovered that there is a similarity between the results of the two tasks (NC vs. eMCI and NC vs. LMCI), many regions are recognized as abnormal in the left hemisphere. These abnormal regions are mainly located in the frontal lobe, temporal lobe and parietal lobe, of which the features of the temporal lobe are more important. For the third task (eMCI vs. LMCI), it can be observed that the most discriminative features of the middle layer are in the right hemisphere, and are mainly located in the frontal lobe and occipital lobe.

Similarly, comparing the results of the ABIDE dataset, it can be seen that many areas are also recognized as abnormal. For the Leuven data, the abnormal regions are mainly located in the frontal lobe, parietal lobe and insula, of which the frontal lobe and insula features are more important. For the NYU data, the abnormal areas are principally situated in the frontal lobe. For the UCLA data, the abnormal areas are mainly located in the frontal lobe, temporal lobe, parietal lobe and occipital lobe, of which the features of the temporal lobe are more important. For the UM data, the abnormal regions are mainly located in the frontal lobe, parietal lobe and insula, of which the features of the frontal lobe are more important. For the USM data, the abnormal areas are mainly located in the frontal lobe, temporal lobe and insula, of which the features of the frontal lobe are of greater importance. It can be noted that most of the tasks captured abnormal information in the frontal lobe and insula.

### C. Visualization of Dementia-Related Features

By encoding and reconstruction, Deep CVAE can obtain better consistency in the data distribution. The following figure gives the regions with the largest difference between the original data and the reconstructed data, which we believe may play a crucial role in the classification results.

Fig. 5 gives the visualization results for the three tasks on the ADNI dataset, where the red area has a stronger effect on the classification results. Comparing the visualization results of these three tasks, it can be found that the left occipital lobe plays an important role in these classification tasks. These abnormal areas are mainly located in the frontal lobe, temporal lobe, and occipital lobe. More precisely, we believe that the

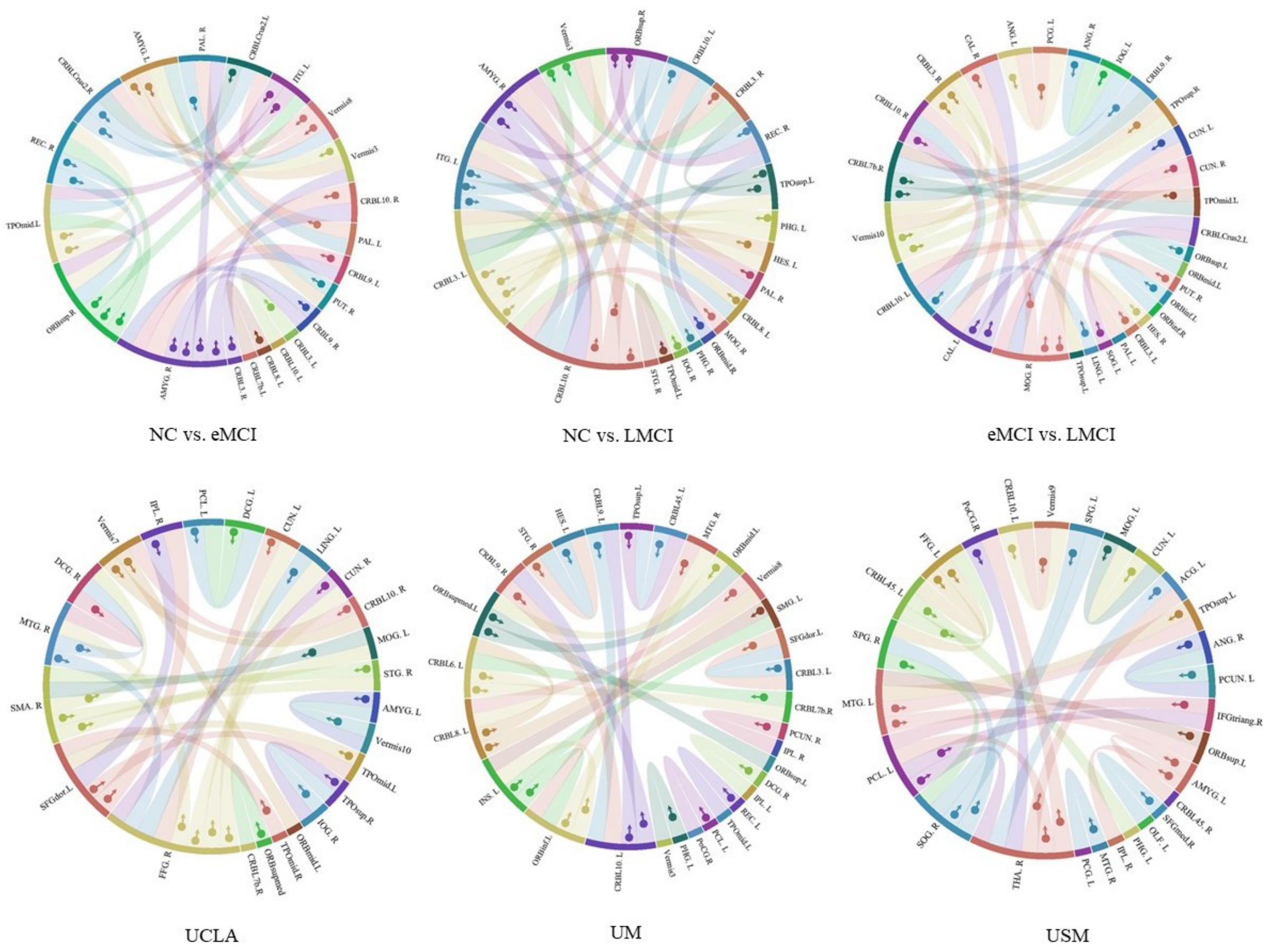


Fig. 3. Visualization results of the most discriminative causal patterns among the six tasks.

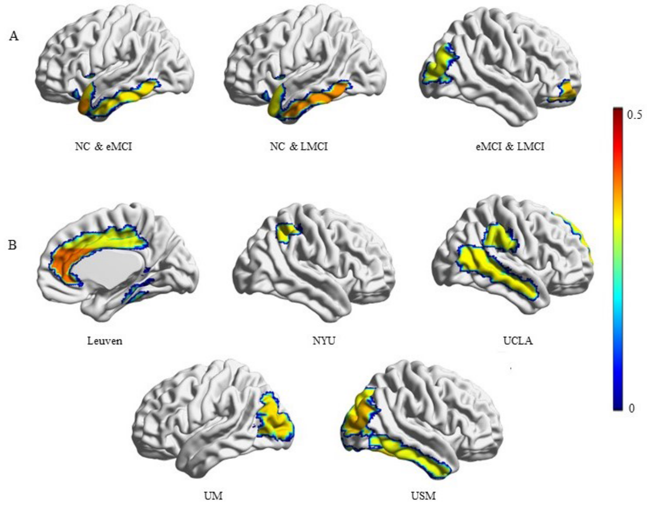
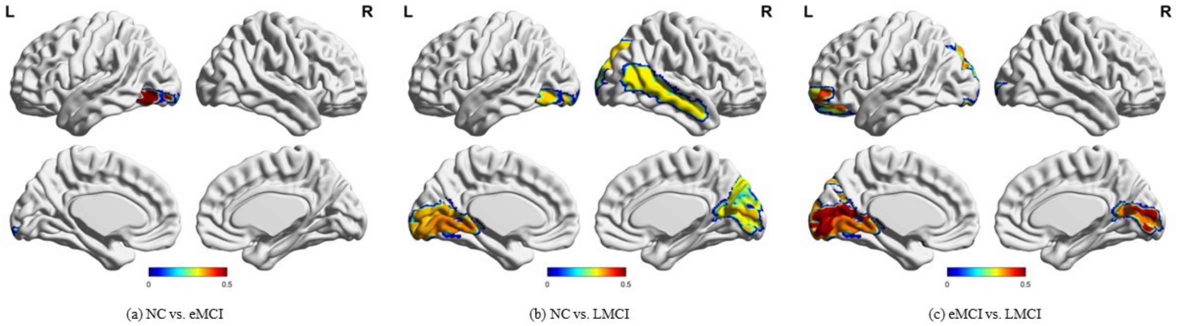


Fig. 4. The hierarchical relationship learned by middle layers of deep CVAE (inference model). (A) for ADNI tasks and (B) for ABIDE tasks.

occipital lobe may contain many biomarkers associated with Alzheimer's disease. The findings of this study are consistent with the findings of several previous studies [11], [41], [42], which indicate the good interpretability of our method.

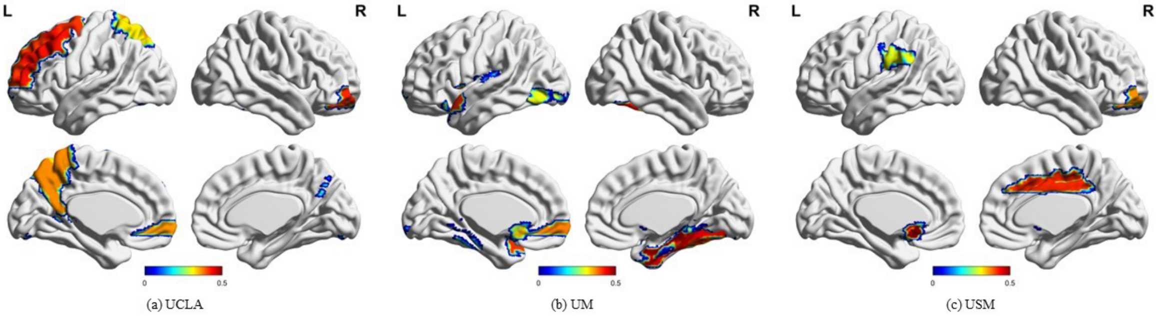
### D. Visualization of ASD-Related Features

As shown in Fig. 6, we present the visualization results of the most discriminative features with ASD on the three datasets. Those areas closer to red are more helpful in improving classification accuracy, while those areas closer to blue are less helpful in improving classification accuracy. Comparing these three results, it can be found that they all have a larger area of abnormal region in the left hemisphere than in the right hemisphere. These abnormal areas are mainly located in the frontal lobe, parietal lobe and insula, among which the frontal lobe seems to be more important. This result is consistent with many previous findings. For example, Cao et al. [43] conducted a classification study of autism using the ABIDE dataset and found that many functional connections between the frontal lobe and the insula were abnormal. Crucitti et al. [44] conducted a cohort study of the frontal lobe in the autistic and non-autistic groups and found significant differences between the frontal lobes of the two groups at the age of 2-4. In addition, we also noticed that the most discriminative features identified by the three are quite different. The specific reasons for this can be summarized as follows. First, ABIDE is a multi-site dataset, and there is domain shift between the data of different sites [40]. Secondly, the differences in data acquisition equipment, parameter settings, and personnel



(a) NC vs. eMCI (b) NC vs. LMCI (c) eMCI vs. LMCI

Fig. 5. Dementia-related brain maps. (a) NC vs. eMCI. (b) NC vs. LMCI. (c) eMCI vs. LMCI.



(a) UCLA (b) UM (c) USM

Fig. 6. ASD-related brain maps. (a) Leuven site. (b) USM site. (c) UCLA site.

operating habits among different sites also have a certain impact on the results.

## V. Conclusion

The discovery of causality is helpful in most disciplines, especially in neuroscience and biology. In this paper, we design a novel framework deep causal variational autoencoder (deep CVAE) for estimating causal effects between brain regions, which contains a causal layer for inferring causal relationships between variables. The core point is the network structure design, which incorporates the solution target as parameters to be learned into the network. The entire network is trained in an end-to-end manner, and it yields excellent performance on two public databases. The experimental results demonstrate the superiority of the method in this paper. In addition to superior performance, the proposed approach offers a new technique for determining causal relationships among a large number of nodes. However, it should be noted that the algorithm employed in this study does not consider the time delay of signal transmission across brain regions, which may introduce biases in estimating causal effects. In our future work, we aim to improve the algorithm by incorporating the estimation of both the causal effects between brain regions and the time delay of signal transmission across them. This refinement will further advance our understanding of the complex dynamics within the brain.

## VI. Declaration

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] J. Peng, P. Wang, N. Zhou, and J. Zhu, “Partial correlation estimation by joint sparse regression models,” *J. Amer. Stat. Assoc.*, vol. 104, no. 486, pp. 735–746, Jun. 2009.

[2] Y. Zhong et al., “Altered effective connectivity patterns of the default mode network in Alzheimer’s disease: An fMRI study,” *Neurosci. Lett.*, vol. 578, pp. 171–175, Aug. 2014.

[3] J. Cai et al., “Dynamic graph theoretical analysis of functional connectivity in Parkinson’s disease: The importance of Fiedler value,” *IEEE J. Biomed. Health Informat.*, vol. 23, no. 4, pp. 1720–1729, Jul. 2019.

[4] J. F. Agastinose Ronicko, J. Thomas, P. Thangavel, V. Koneru, G. Langs, and J. Dauwels, “Diagnostic classification of autism using resting-state fMRI data improves with full correlation functional brain connectivity compared to partial correlation,” *J. Neurosci. Methods*, vol. 345, Nov. 2020, Art. no. 108884.

[5] N. Leonardi et al., “Principal components of functional connectivity: A new approach to study dynamic brain connectivity during rest,” *NeuroImage*, vol. 83, pp. 937–950, Dec. 2013.

[6] E. Damaraju et al., “Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia,” *NeuroImage, Clin.*, vol. 5, pp. 298–308, Jan. 2014.

[7] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, “Nonlinear causal discovery with additive noise models,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 21, 2008, pp. 1–8.

[8] K. Harada and H. Fujisawa, “Sparse estimation of linear non-Gaussian acyclic model for causal discovery,” *Neurocomputing*, vol. 459, pp. 223–233, Oct. 2021.

[9] W. Liao et al., “Altered effective connectivity network of the amygdala in social anxiety disorder: A resting-state fMRI study,” *PLoS ONE*, vol. 5, no. 12, Dec. 2010, Art. no. e15238.

[10] S. H. Hojjati, A. Ebrahimzadeh, A. Khazaee, and A. Babajani-Feremi, “Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM,” *J. Neurosci. Methods*, vol. 282, pp. 69–80, Apr. 2017.

[11] L. Qiao, H. Zhang, M. Kim, S. Teng, L. Zhang, and D. Shen, “Estimating functional brain networks by incorporating a modularity prior,” *NeuroImage*, vol. 141, pp. 399–407, Nov. 2016.

[12] Z. Wang, J. Xin, Z. Wang, Y. Yao, Y. Zhao, and W. Qian, “Brain functional network modeling and analysis based on fMRI: A systematic review,” *Cognit. Neurodynamics*, vol. 15, no. 3, pp. 389–403, Jun. 2021.

[13] C. Berzuini, P. Dawid, and L. Bernardinell, *Causality: Statistical Perspectives and Applications*. Hoboken, NJ, USA: Wiley, 2012.

[14] M. Nauta, D. Bucur, and C. Seifert, “Causal discovery with attention-based convolutional neural networks,” *Mach. Learn. Knowl. Extraction*, vol. 1, no. 1, pp. 312–340, Jan. 2019.

[15] Z. Duan, H. Xu, Y. Huang, J. Feng, and Y. Wang, “Multivariate time series forecasting with transfer entropy graph,” *Tsinghua Sci. Technol.*, vol. 28, no. 1, pp. 141–149, Feb. 2023.

[16] S. Wein et al., “A graph neural network framework for causal inference in brain networks,” *Sci. Rep.*, vol. 11, p. 8061, Apr. 2021.

[17] J. Liu, J. Ji, X. Jia, and A. Zhang, “Learning brain effective connectivity network structure using ant colony optimization combining with voxel activation information,” *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 2028–2040, Jul. 2020.

[18] L. Zhou, L. Wang, L. Liu, P. Ogunbona, and D. Shen, “Learning discriminative Bayesian networks from high-dimensional continuous neuroimaging data,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2269–2283, Nov. 2016.

[19] Y. Yu, J. Chen, T. Gao, and M. Yu, “DAG-GNN: DAG structure learning with graph neural networks,” in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 7154–7163.

[20] D. Kalainathan, O. Goudet, I. Guyon, D. Lopez-Paz, and M. Sebag, “Structural agnostic modeling: Adversarial learning of causal graphs,” *J. Mach. Learn. Res.*, vol. 23, no. 1, pp. 9831–9892, 2022.

[21] J. Liu, J. Ji, G. Xun, L. Yao, M. Huai, and A. Zhang, “EC-GAN: Inferring brain effective connectivity via generative adversarial networks,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4852–4859.

[22] G. Zhou, L. Yao, X. Xu, C. Wang, L. Zhu, and K. Zhang, “On the opportunity of causal deep generative models: A survey and future directions,” 2023, *arXiv:2301.12351*.

[23] N. T. Markov et al., “A weighted and directed interareal connectivity matrix for macaque cerebral cortex,” *Cerebral Cortex*, vol. 24, no. 1, pp. 17–36, Jan. 2014.

[24] G. Z. Tau and B. S. Peterson, “Normal development of brain circuits,” *Neuropsychopharmacology*, vol. 35, no. 1, pp. 147–168, Jan. 2010.

[25] S. Bang et al., “Engineered neural circuits for modeling brain physiology and neuropathology,” *Acta Biomaterialia*, vol. 132, pp. 379–400, Sep. 2021.

[26] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, “Spectral networks and locally connected networks on graphs,” 2013, *arXiv:1312.6203*.

[27] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” 2016, *arXiv:1609.02907*.

[28] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[29] I. Ng, S. Zhu, Z. Fang, H. Li, Z. Chen, and J. Wang, “Masked gradient-based causal structure learning,” in *Proc. 2022 SIAM Int. Conf. Data Mining (SDM)*, 2022, pp. 424–432.

[30] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” 2013, *arXiv:1312.6114*.

[31] S. L. Bressler and A. K. Seth, “Wiener–Granger causality: A well established methodology,” *NeuroImage*, vol. 58, no. 2, pp. 323–329, Sep. 2011.

[32] F. Parente and A. Colosimo, “Modelling a multiplex brain network by local transfer entropy,” *Sci. Rep.*, vol. 11, no. 1, p. 15525, Jul. 2021.

[33] Z. Xia, T. Zhou, S. Mamoon, A. Alfakih, and J. Lu, “A structure-guided effective and temporal-lag connectivity network for revealing brain disorder mechanisms,” *IEEE J. Biomed. Health Informat.*, vol. 27, no. 6, pp. 2990–3001, Apr. 2023.

[34] H. Li, S. Yu, and J. Principe, “Causal recurrent variational autoencoder for medical time series generation,” 2023, *arXiv:2301.06574*.

[35] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.

[36] J. Franklin, “The elements of statistical learning: Data mining, inference and prediction,” *Math. Intelligencer*, vol. 27, no. 2, pp. 83–85, Mar. 2005.

[37] C. Echávarri et al., “Atrophy in the parahippocampal gyrus as an early biomarker of Alzheimer’s disease,” *Brain Struct. Function*, vol. 215, nos. 3–4, pp. 265–271, Jan. 2011.

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2017, pp. 618–626.

[39] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” 2014, *arXiv:1412.6806*.

[40] M. Wang, D. Zhang, J. Huang, P.-T. Yap, D. Shen, and M. Liu, “Identifying autism spectrum disorder with multi-site fMRI via low-rank domain adaptation,” *IEEE Trans. Med. Imag.*, vol. 39, no. 3, pp. 644–655, Mar. 2020.

[41] X. Chen, H. Zhang, Y. Gao, C. Wee, G. Li, and D. Shen, “High-order resting-state functional connectivity network for MCI classification,” *Hum. Brain Mapping*, vol. 37, no. 9, pp. 3282–3296, Sep. 2016.

[42] T. Zhou, K.-H. Thung, M. Liu, F. Shi, C. Zhang, and D. Shen, “Multi-modal latent space inducing ensemble SVM classifier for early dementia diagnosis with neuroimaging data,” *Med. Image Anal.*, vol. 60, Feb. 2020, Art. no. 101630.

[43] P. Cao, G. Wen, X. Liu, J. Yang, and O. R. Zaiane, “Modeling the dynamic brain network representation for autism spectrum disorder diagnosis,” *Med. Biol. Eng. Comput.*, vol. 60, no. 7, pp. 1897–1913, Jul. 2022.

[44] J. Crucitti, C. Hyde, P. G. Enticott, and M. A. Stokes, “A systematic review of frontal lobe volume in autism spectrum disorder revealing distinct trajectories,” *J. Integrative Neurosci.*, vol. 21, no. 2, p. 57, 2022.