# Deep Unsupervised Representation Learning for Feature-Informed EEG Domain Extraction

Han Wei Ng, *Member, IEEE*, and Cuntai Guan, *Fellow, IEEE*

***Abstract*— In electroencephalography (EEG) classification paradigms, data from a target subject is often difficult to obtain, leading to difficulties in training a robust deep learning network. Transfer learning and their variations are effective tools in improving such models suffering from lack of data. However, many of the proposed variations and deep models often rely on a single assumed distribution to represent the latent features which may not scale well due to inter- and intra-subject variations in signals. This leads to significant instability in individual subject decoding performances. The presence of non-trivial domain differences between different sets of training or transfer learning data causes poorer model generalization towards the target subject. However, the detection of these domain differences is often difficult to perform due to the ill-defined nature of the EEG domain features. This study proposes a novel inference model, the Joint Embedding Variational Autoencoder, that offers conditionally tighter approximation of the estimated spatiotemporal feature distribution through the use of jointly optimised variational autoencoders to achieve optimizable data dependent inputs as an additional variable for improved overall model optimisation and scaling without sacrificing model tightness. To learn the variational bound, we show that maximising the marginal log-likelihood of only the second embedding section is required to achieve conditionally tighter lower bounds. Furthermore, we show that this model provides state-of-the-art EEG data reconstruction and deep feature extraction. The extracted domains of the EEG signals across each subject displays the rationale as to why there exists disparity between subjects' adaptation efficacy.**

***Index Terms*— Deep representation learning, feature extraction, signal processing, spatiotemporal data.**

## I. Introduction

TO OVERCOME the lack of data and poor subject-specific performance in electroencephalography (EEG) classification tasks, fine-tuning techniques such as transfer-learning [1] are often applied. However, when conducting further fine-tuning [2] or adaptation [3], the change in decoding accuracy across different subjects is often non-uniform [1]. Furthermore, improvement in classification accuracy due to adaptation in offline studies [4] does not necessarily translate towards online decoding scenarios [3]. Adaptation techniques that focus on subject-specific adaptation [5], [6] do not fully leverage upon the data from other subjects which may be beneficial towards improving decoding accuracy. Although the overall decoding accuracy may increase with such fine-tuning techniques, closer observations on individual subject performances show that a number of individuals experience a large decrease in classification accuracy which plagues modern neural networks [1] resulting in the difficulty in implementing such models for real-world general usage.

A reason for the instability in the performance across subjects may be due to the presence of ill-defined non-stationary features that vary from subject to subject [7]. Although subject-specific transfer-learning [1], [5], [6] may be utilized to overcome this instability by fine-tuning using the target subject's data, this is based on the assumption that the adaptation data comes from the same continuous distribution from the target evaluation data. However, this assumption may not hold true across time due to the physiological changes of the neural signals across time [8], resulting in intra-subject variability [9]. In addition, recent studies showed the importance of taking into account the differences across different sessions for each individual subject [10]. Using the proposed methodology in this work, we further extend on representation learning to reveal that even within subjects, there exists a non-trivial intra-subject non-stationary domain differences (Fig. 9, 4) which are associated with the variations in EEG recording phases.

In earlier methods such as the filter bank common spatial pattern (FBCSP) [11], the EEG signals are processed by splitting the signal into separate bands corresponding to different frequencies of the brain signals. The network is then able to learn and pick up on the important features dependent on each of the bands. In more recent works, this is further expanded later in Filter-Bank Convolutional Networks (FBCNet) [12] where multi-view data representation is utilized followed by spatial filtering to extract spectro-spatially discriminative features to train the network. These methods lay the foundation of extracting vital spatial-temporal features of EEG data for classification tasks, but still hold the disadvantage of not taking

Manuscript received 27 August 2023; revised 1 November 2023; accepted 29 November 2023. Date of publication 4 December 2023; date of current version 14 December 2023. This work was supported in part by the National Research Foundation, Singapore under its AI Singapore Programme (AISG), under Award AISG2-PhD-2021-08-021; and in part by the Research Innovation Enterprise 2020 Advanced Manufacturing and Engineering (RIE2020 AME) Programmatic Fund, Singapore, under Grant A20G8b0102. *(Corresponding author: Han Wei Ng.)*

Han Wei Ng is with the Ph.D. Fellowship Programme, AI Singapore, Singapore 117602, and also with the Faculty of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: hanwei001@e.ntu.edu.sg).

Cuntai Guan is with the Faculty of Computer Science and Engineering, Nanyang Technological University, Singapore 639798.

Digital Object Identifier 10.1109/TNSRE.2023.3339179

into consideration the temporal variation within-subject across time [8] which may lead to poorer outcomes in network classification.

Representation learning [13] is an important area of research that breaks down a given input into a set of features that best represent the original data. When designing networks, the number of features is usually predetermined [13], [14], [15] and in some cases specifying the nature of the discriminative features can lead to better performance [14]. One such network used is the Variational Autoencoder (VAE) [15], which is used to break input data into encoded features that best represent the original data. Features play a vital role towards providing better understanding of the underlying common attributes between the signals. By doing so, techniques such as domain generalisation [16], [17] and/or adaptation [18], [19] can be applied to achieve robust training and performance in deep neural networks. Although previous works on motor imagery datasets [20] employ fuzzy representation learning methods of comparing similarity and dissimilarity between sets of interval-valued EEG data, the performance is still highly limited due to its restrictive architecture.

Thus far, fewer work has been done on representation learning on spatiotemporal signals [21] compared to more well established datasets such as images [13], [22], [23] or natural language [23], [24]. This is especially so for feature extraction of common biophysical signals such as electroencephalography (EEG), electromyography (EMG) and electrocardiography (ECG). In the case of EEG signals, common methods in training subject-independent models of EEG classifiers often face significant variations in subject-to-subject performance.

Presently, traditional VAEs extracts features based on the assumption that the best representation of the determined latent features of any given data all falls under a similar or identical distribution. Additionally, the VAE assumes that the given model structure is best able to generalise and encode all the latent spatiotemporal features concurrently. Given the complexity of domains associated with spatiotemporal data [25], there exist multiple possible paradigms that the features encode for. In this study, we discover that one common embedding model structure may not be ideal in computing overall meaningful latent features.

Therefore, this study aims to provide an adaptive model framework that leverages upon both subject-independent and subject-specific data, while taking into consideration the intra- and inter-subject variability of EEG signals. The main motivation behind this study is to construct a domain-agnostic methodology that enables unsupervised detection of domain differences in EEG signals via hidden features. This is vital towards building robust networks for lifelong learning via applications of representation learning to perform data selection as well as one/few-shot adaptation towards unseen test subject data that may have significantly different underlying features than the training dataset.

The contributions of the study are as follows. (1) We propose a new framework based on sequential joint embedding sections of VAEs utilising encoder-decoder pair networks. (2) This new framework allows the splitting of latent features into separate posterior distributions which enables better representation of the true distribution. (3) In addition, this framework further increases the flexibility of the model to allow separate model structures to compute each separate set of latent features. (4) Finally, a new loss function is introduced which offers significantly better feature learning and reconstruction accuracy through modelling the residual loss of the network using a conditionally tighter lower bound on the true log-likelihood.

## II. Background: Variational Autoencoders

Traditional VAEs consist of probabilistic encoder-decoder pairs. For a given input $x$, the encoder is an inference model with weights and biases $\theta$ which gives the hidden latent variables as output $z$. The inference model is thus given by $q_\Theta(z|x)$, a Gaussian probability distribution. For the same VAE, a decoder model with weights and biases $\phi$ is given by a joint probability $p_\Phi(x,z) = p_\Phi(x|z)p(z)$. During training, the encoder and decoders are trained simultaneously by finding the parameters that best optimise the variational lower bound of the likelihood $p_\phi(x) = \int p_\phi(x,z)dz$.

Thus, the effectiveness of the VAE in reconstructing the original input is given by the reconstruction log-likelihood $logp_\Phi(x|z)$. The reconstruction loss function is therefore given by the expected negative log-likelihood $-\mathbb{E}_{q_\Theta(z|x)}[\log\ p_\Phi(x|z)]$ computed with respect to the distribution of the latent features by the encoder.

In addition to the reconstruction loss, VAEs also take into account a regularisation term given by the Kullback-Leibler (KL) divergence between two continuous distributions [26], the encoder's variational posterior $q_\Theta(z|x)$ and the prior $p(z)$ where the latent variables are sampled from. The divergence measures how close the two distributions $q$ and $p$ are to each other and is given by $\mathbb{KL}(q_\Theta(z|x)||p(z))$. Therefore, the overall loss function $L_i$ for an input datapoint $x_i$ is:

$$L_i = -\mathbb{E}_{q_\Theta(z|x_i)}[\log\ p_\Phi(x_i|z)] + \mathbb{KL}(q_\Theta(z|x_i)||p(z)) \quad (1)$$

The VAE loss function relies on mathematical convex optimisation principles to ensure that there indeed exists an evidence lower bound (ELBO) [27]. Logarithm functions are strictly concave, since the negative of concave functions are convex, the negative log-likelihood is convex. Expectation preserves the convexity of the function and thus the expected negative log-likelihood is a convex function. Using Jensen's inequality [28], we can also prove that the KL divergence is a convex function with respect to a pair of probability distributions $p$ and $q$.

## III. Joint Embedding Variational Autoencoder

### A. Model Overview

We propose a new model as shown in Figure 1 consisting of two jointly connected encoder-decoder networks via a new loss function that offers a tighter lower bound when certain conditions are enforced. The intuition behind the proposed framework lies in computing the residual loss of the initial part of the network before systematically recovering the loss. The first encoder-decoder pair is constructed identically to a traditional VAE which takes in an input $x$ and gives a

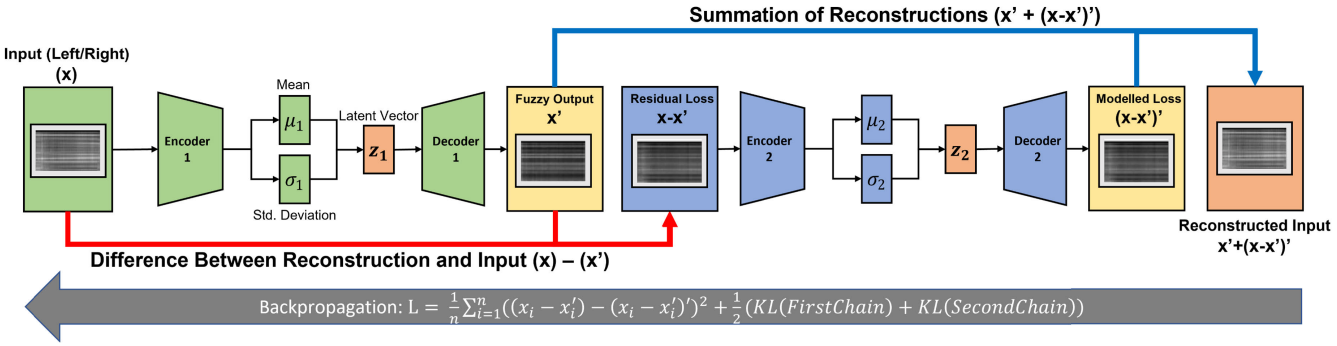


Fig. 1. Overall Joint Embedding Variational Autoencoder framework comprising of two connected encoder-decoder pairs. Takes in an input $x$ and gives a reconstructed output $x' + (x - x')'$. Note that the overall reconstruction accuracy of the input depends solely on the second embedding section's ability to reconstruct the residual reconstruction difference. Learnt joint embeddings $z_1$ and $z_2$ allow for separate dependent distributions to be utilized to model the latent features, enabling greater flexibility and accuracy as compared to a single distribution. Generation capabilities are retained while maximizing reconstruction accuracy.

noisy output $x'$. We can form an equation for the discrepancy between the original input and reconstructed output by the initial VAE as $x - x'$:

Thus, the network embedding section of the JEVAE attempts to construct the probability distribution for $x$ with respect to latent variables:

$$p_\phi(z_1) = p_\phi(z_{F_{z1}}) \prod_{i=1}^{F_{z1}-1} p_\phi(z_i|z_{i+1}),$$
$$p_\phi(z_2) = p_\phi(z_{F_{z2}}) \prod_{j=1}^{F_{z2}-1} p_\phi(z_j|z_{j+1}) \quad (2)$$

where $F_{z1}$, $F_{z2}$ represents the dimensionality of the latent space in the first and second network embedding section respectively.

The second embedding section in JEVAE takes in the discrepancy between the input to the previous decoder and the reconstructed output that is based on the computation of the intial VAE. This part of the framework thus aims to reconstruct the information loss from the first VAE by taking in the input $x - x'$ and giving an approximate output $(x - x')'$.

The expected negative log-likelihood of only the second network with respect to the first network is:

$$\log p(x - x') = \mathbb{E}_{q_\Theta(z_2|x-x')}[\log\ p_\Phi(x - x'|z_2)] - \mathbb{KL}(q_\Theta(z_2|x - x')||p(z_2)) \quad (3)$$

In this study, we mainly study biophysical signals such as EEG which are well represented via Guassian distributions [29]. Thus, assuming a continuous Gaussian distribution for both latent spaces, each of the random hidden variables sets $z_1$ and $z_2$ can be individually expressed as:

$$p_\phi(z_{Ri}) = \mathcal{N}(z_{Ri}|0, \mathrm{I}), \quad p_\phi(z_i|z_{i+1}) = \mathcal{N}(z_i|\mu_{i,p}(z_{i+1}), \sigma^2_{i,p}(z_{i+1})) \ \text{ for i} = 1, 2 \quad (4)$$

Since the second network aims to encode the information that is not encoded in the initial network, the overall reconstruction of the original input is now given by the sum of both of the outputs of the two networks $x' + (x - x')'$. Importantly, both the first and second VAEs are trained simultaneously

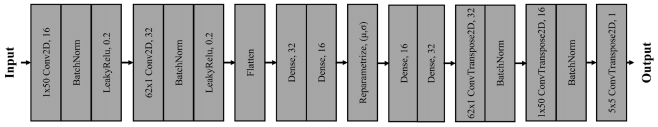


Fig. 2. Overall VAE network architecture. The network architecture utilizes spatial and temporal-level filters to identify latent features that best represent the input EEG signals. The VAE network is sequentially connected twice via the proposed method to achieve an end-to-end dual-VAE network capable of extracting joint embeddings.

with both parts of the framework being linked via the input to the second network which depends on the output of the first network, resembling a embedding section. Thus the overall neural network is denoted as Joint Embedding Variational Encoder (JEVAE).

Although there exists variations of the VAE such as the Ladder VAE (LVAE) [30] that utilise the separation of latent features, where the JEVAE differs is in how the subsequent set of latent features is derived. Hierarchical methods rely on directly applying a consecutive simple decoder or a series of decoders [31], [32], [33] on the previous decoder output to achieve deeper extraction of features. This is in direct contrast to the horizontal method presented here whereby the input to the subsequent decoder depends not only on the previous decoder's output but also the original input to the previous decoder. Additionally, most of these state-of-the-art VAEs are built for image sets and often do not support non-regular datasets with different height to widths or take into account the existence of temporal features within each image.

The architectural design of the VAE (Fig. 2) draws inspiration from the highly effective DeepConvNet utilized in motor imagery classification as outlined by Schirrmeister et al. [34]. The robust classification performance achieved by this architecture suggests its ability to discern critical distinguishing features across various target classes. This capability is facilitated by employing distinct temporal convolutions succeeded by spatial filters. In a similar vein, our proposed VAE configuration incorporates temporal convolutions followed by spatial filters, ultimately leading to effective dimensional reduction. Consequently, the VAE

demonstrates proficiency in minimizing reconstruction errors and capturing essential latent features.

In summary, the study explores the concept of jointly separated latent variables which are split into two sets, each having its own probability distribution. The two probability distributions of the hidden representations are connected to one another through the optimisation function given by the maximisation of the variational lower bound.

### B. Variational Lower Bound

To achieve better model generalisation, the KL loss is applied to both embedding sections separately. The two sets of features are represented as two separate individual distributions. For the proposed framework, there exists multiple possible loss functions that can optimise the overall network. For instance, the most straightforward method would be to set the loss function as the sum of the two KL divergence losses and the individual network reconstruction loss.

$$\begin{aligned} ELBO = &\mathbb{E}_{q_\Theta(z_1|x)}[\log\ p_\Phi(x|z_1)] \\ &- \mathbb{KL}(q_\Theta(z_1|x)||p(z_1)) \\ &+ \mathbb{E}_{q_\Theta(z_2|x-x')}[\log\ p_\Phi(x-x'|z_2)] \\ &- \mathbb{KL}(q_\Theta(z_2|x-x')||p(z_2)) \end{aligned} \tag{5}$$

Next, we further examine the reconstruction loss of the JEVAE network. Equivalently, since we assume that the latent space vectors follow a gaussian distribution, as the authors of the original VAE [15] highlighted, the decoding term $\log\ p_\Phi(x|z)$ is Gaussian Multi-Layer Perceptron (MLP) [35]. Thus, the maximising of the reconstruction log-likelihood can also be achieved via the minimising of the mean squared error (MSE) between the original input and the predicted output of the decoder network.

Although the above equation 5 does have a variational lower bound for the overall JEVAE network, it is not the most ideal case. Consider an alternative loss function involving the sum of the two KL divergence losses and the overall network reconstruction loss. The overall reconstruction loss would be given by $\mathbb{E}_{q_\Theta(z_1,z_2|x,x-x')}[\log\ p_\Phi(x|z_1,z_2)]$ and ELBO of the log-likelihood:

$$\begin{aligned} \log p(x) \geq ELBO = &\mathbb{E}_{q_\Theta(z_1,z_2|x,x-x')}[\log\ p_\Phi(x|z_1,z_2)] \\ &- \mathbb{KL}(q_\Theta(z_1|x)||p(z_1)) \\ &- \mathbb{KL}(q_\Theta(z_2|x-x')||p(z_2)) \end{aligned} \tag{6}$$

Assuming a Gaussian distribution for both latent variables $z_1$ and $z_2$, the mean squared error between the input and the overall reconstructed output of the JEVAE is:

$$ReconLoss_{\text{JEVAE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - (x_i' + (x_i - x_i')'))^2 \tag{7}$$

where n is the total number of datapoints, $x_i$ is the input for the overall JEVAE model and
$x_i' + (x_i - x_i')'$ is the reconstructed output.

Here, we prove that the overall network reconstruction loss is therefore identical to the reconstruction loss of solely the second network. The loss of the second network in the JEVAE is given by:

$$ReconLoss_2 = \frac{1}{n}\sum_{i=1}^{n}((x_i - x_i') - (x_i - x_i')')^2 \tag{8}$$

where $x_i - x_i'$ is the input to the second network embedding section and $(x_i - x_i')'$ is the reconstruction of the second network input.

By expanding Equation 7:

$$\begin{aligned} ReconLoss_{\text{JEVAE}} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - x_i' - (x_i - x_i')')^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}((x_i - x_i') - (x_i - x_i')')^2 \\ &== ReconLoss_2 \end{aligned} \tag{9}$$

Therefore, maximising the overall log-likelihood $\mathbb{E}_{q_\Theta(z_1,z_2|x,x-x')}[\log\ p_\Phi(x|z_1,z_2)]$ is equvilant to maximising the log-likelihood of the second embedding section $\mathbb{E}_{q_\Theta(z_2|x-x')}[\log\ p_\Phi(x - x'|z_2)]$. When considering the computation of the KL divergence for JEVAE, the summation of two KL losses is likely to lead to an increase in the weightage towards KL loss in JEVAE. This is due to how KL divergence is calculated. Since the loss is computed such that $\sum_{x\epsilon\mathcal{X}} q_\Theta(z|x_i) = 1$, by including a summation of two separate KL divergences, the possible range of the loss is scaled by a factor of 2. Therefore, we propose that the loss in JEVAE should be scaled back to prevent uneven weightage towards the KL loss. This can be achieved by keeping $\sum_{x\epsilon\mathcal{X}} q_\Theta(z_1|x) + \sum_{x\epsilon\mathcal{X}} q_\Theta(z_2|x - x') = 1$. This can be done by performing a similar implementation to Beta-VAEs [36] whereby an adjustable hyperparameter is introduced to balance the latent channel capacity and independence constraints with the expected log-likelihood. In the case of JEVAE, since there are two KL losses, we introduce two hyperparameters $\alpha$ and $\beta$. Although these hyperparameters can be adjusted, we introduce a formulation for the baseline of these hyperparameters.

The variational lower bound function of the JEVAE can therefore be simplified to:

$$\begin{aligned} ELBO = &\mathbb{E}_{q_\Theta(z_2|x-x')}[\log\ p_\Phi(x - x'|z_2)] \\ &- \alpha\mathbb{KL}(q_\Theta(z_1|x)||p(z_1)) \\ &- \beta\mathbb{KL}(q_\Theta(z_2|x-x')||p(z_2)) \end{aligned} \tag{10}$$

where given the number of features in the first embedding section, $F_{z1}$ and the second embedding section, $F_{z2}$:

$$\alpha = \frac{F_{z1}}{(F_{z1} + F_{z2})},\ \beta = \frac{F_{z2}}{(F_{z1} + F_{z2})} \tag{11}$$

Intuitively, we can observe that in the case whereby the second embedding section is a perfect autoencoder, where $(x - x')' == x - x'$. The reconstructed input by JEVAE is obtained via the summation of the outputs of the individual sections, $x' + (x - x')'$. Since $(x - x')' == x - x'$, the reconstruction is now given by $x' + (x - x')$ which is equivalent to the original input $x$.

### C. Optimisation Tightness of Overall Network

Next, we show that the JEVAE framework offers a conditionally tighter ELBO. Through variational inference, a tractable lower bound can be computed on the log-likelihood which is also used as the training loss function as shown in equation 1. For a traditional VAE, the ELBO on log-likelihood $\log p(x)$ is given by:

$$\log p(x) \geq \mathbb{E}_{q_\Theta(z|x)}[\log\ p_\Phi(x|z)] - \mathbb{KL}(q_\Theta(z|x)||p(z)) \quad (12)$$

Since the performance of the JEVAE model depends entirely on the second network's reconstruction as shown in equation 7, from the JEVAE ELBO equation 10, the new ELBO on JEVAE on the log-likelihood is:

$$\begin{aligned}\log p(x-x') \geq\ & \mathbb{E}_{q_\Theta(z_2|x-x')}[\log\ p_\Phi(x-x'|z_2)] \\ & - \alpha\mathbb{KL}(q_\Theta(z_1|x)||p(z_1)) \\ & - \beta\mathbb{KL}(q_\Theta(z_2|x-x')||p(z_2)) \end{aligned} \quad (13)$$

This differs from vanilla VAEs due to the difference in the log-likelihood terms. Thus, the computation of the tightness of the lower bound is different as well. The variational lower bound for JEVAE is conditionally tighter than traditional VAEs, subject to the following conditions:

$$\begin{aligned}&\log p(x-x') - \mathbb{E}_{q_\Theta(z_2|x-x')}[\log\ p_\Phi(x-x'|z_2)] \\ &\quad \leq \log p(x) - \mathbb{E}_{q_\Theta(z|x)}[\log\ p_\Phi(x|z)] \end{aligned} \quad (14)$$

$$\begin{aligned}&\alpha\mathbb{KL}(q_\Theta(z_1|x)||p(z_1)) + \beta\mathbb{KL}(q_\Theta(z_2|x-x')||p(z_2)) \\ &\quad \leq \mathbb{KL}(q_\Theta(z|x)||p(z)) \end{aligned} \quad (15)$$

The reason why the proposed JEVAE has conditionally tighter lower bounds is due to the conditional equation 14 being non-guaranteed in nature. For equation 14, since the number of latent features used to represent the data is reduced to half, the second network embedding section is less able to capture as much information compared to using a larger feature size. However, the second network aims to approximate the log-likelihood $\log p(x-x')$, which has the variable input $x'$. In the case whereby $x' > 0$, the log-likelihood $\log p(x-x')$ will be less than $\log p(x)$ due to the strictly increasing nature of logarithmic functions. This is beneficial since reduction in the actual log-likelihood would reduce the difference from the approximated log-likelihood. But in the case where $x' < 0$, $\log p(x-x')$ becomes greater than $\log p(x)$, causing the difference to increase.

Finally, the degree of tightness depends on how great the difference is between the JEVAE and VAE's log-likelihood and expectation for 14. One method to ensure equation 14 is to set the number of latent variables for the second embedding section $z_2$ to be equal to the number of latent variables for a similar counterpart traditional VAE $z$ and restrict $x' \geq 0$. However, this would result in greater dimensionality in JEVAE compared to traditional VAE which is not desirable. It is important to note having higher ELBOs does not always necessarily indicate a better performing model and vice versa [37].

The reconstruction log-likelihood $\log\ p_\Phi(x-x'|z_2)$ is dependent on $z_2$, and $z_2$ is represented by the Gaussian probability distribution $q_\Theta(z_2|x-x')$, an optimal $x'$ probability distribution can be obtained alongside the weights and biases that maximises the reconstruction log-likelihood.

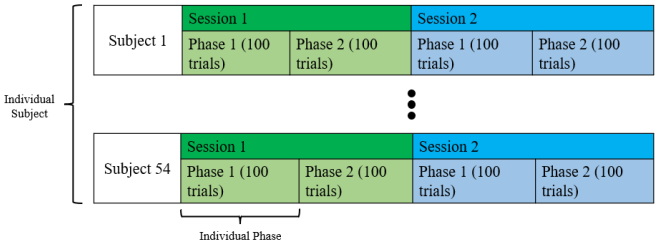


Fig. 3. Overall dataset architecture of 54 subjects conducting bi-class motor imagery.

Immediately, we can observe that one of the benefits of using the JEVAE is the ability to transform the input of the maximisation problem into a variable rather than a static value determined by the known samples. The maximisation of the overall ELBO is therefore entirely dependent on a variable input $x-x'$ rather than a static input $x$ which introduces greater flexibility to the network. Previous work done has shown the value in introducing learnable inputs to VAEs [38]. In addition, this results in only having to optimise for the second set of latent variables $z_2$ which again is dependent on the variable input $x - x'$. This differs from the traditional VAE which relies on the static input $x$ to compute the optimal Gaussian distribution of the latent vectors alongside the posterior and likelihood distributions.

The separability of latent features has previously been shown to benefit feature extraction performance [39], [40]. Thus, the loss function enables the model to split the encoding work between the two embedding sections to find an optimal balance.

## IV. Experimental Setup

### A. Dataset

To evaluate the effectiveness of the proposed JEVAE compared to traditional VAE in the context of multichannel EEG signals, we utilise a well-established EEG dataset by the Department of Brain and Cognitive Engineering, Korea University [41]. The dataset consists of 54 healthy subjects (ages 24–35) performing binary class motor imagery (MI) tasks (Fig. 3), where their EEG signals were recorded using BrainAmp (Brain Products; Munich, Germany) with 62 Ag/AgCl electrodes at a sampling rate of 1000 Hz. All subjects remain anonymous and the design of the experiments follows the protocol set by Pfurtscheller and Neuper [42].

Further details about the data and protocols can be found in [41]. Each subject participated in two data recording sessions across two different days with a total of 400 trials. The first session consists of offline training for recording data and training a baseline classifier, and the second session involves online testing where visual feedback is provided to the subject by decoding live data via the classifier. Each session consists of 200 trials, split into two phases of 100 trials. Pre-processing of the data was done following similar work done on EEG data by Zhang et al. [1].

TABLE I
APPROXIMATED TEST LOG-LIKELIHOOD SCORES FOR MODELS TRAINED ON EEG. LOWER SCORES INDICATE STRONGER MODEL PERFORMANCE IN RECONSTRUCTING INPUT DATA

| Number of Features | VAE | VAE + Clip | JEVAE | $\beta$-JEVAE | $\beta$-JEVAE + Clip |
|---|---|---|---|---|---|
| 8 | -87.2 (±120.94) | -90.1 (±122.10) | -100.8 (±205.32) | -78.6 (±102.46) | -135.2 (±400.39) |
| 16 | -91.2 (±126.00) | -89.0 (±128.84) | -82.5 (±105.43) | **-74.8 (±91.56)** | -84.0 (±117.75) |
| 32 | -104.5 (±130.00) | -87.3 (±122.38) | -90.5 (±117.33) | -78.4 (±98.76) | -88.4 (±103.59) |

TABLE II
APPROXIMATED TEST LOG-LIKELIHOOD SCORES FOR JEVAE USING DIFFERENT LOSS FUNCTIONS. *Default* LOSS FUNCTION IS AS PROPOSED IN EQUATION 10. THE *full* LOSS FUNCTION USES THE CONVENTIONAL INPUT RECONSTRUCTION LOSS. *Indiv* USES THE SUM OF RECONSTRUCTION LOSSES FOR EACH INDIVIDUAL SECTION

| Loss Function | $\beta$-JEVAE - default | $\beta$-JEVAE - full | $\beta$-JEVAE - indiv |
|---|---|---|---|
| JEVAE (16 Features) | **-74.8 (±91.56)** | -84.2 (±110.93) | -101.9 (±173.47) |

### B. Model Training Details

In this study, we focus mainly on EEG signals. EEG signals have both spatial and temporal features associated with the relative electrode positions and the activation patterns of the brain. Thus, when constructing the two network embedding sections for the JEVAE, the JEVAE has both spatial and temporal convolutional layers to take into account the relationships in both domains. After each convolutional layer, batch normalisation [43] is applied. For some of the models, a gradient clipping threshold of 0.25 is used.

All models in this study were trained end-to-end using the Adam [44] optimiser with a mini-batch size of 16, utilising the ELBO computation equation as shown in 10. For each of the 54 subjects, we report the test approximated log-likelihood scores. The EEG data is split into a subject specific 80-10-10 train-validation-test split. The models receive a $62\times1000$ input and were trained for 100 epochs per subject with a learning rate of 0.0005 and a weight decay of 0.00025.

## V. RESULTS

### A. Model Performance: Extraction of EEG Domain

Since the log-likelihood of the overall network is no longer dependent on solely on variable $x$, but is changed to the log-likelihood $\log p(x-x')$. Thus, it would not be a fair comparison to compare the variational ELBOs between the traditional VAE and the proposed JEVAE. However, we still display that the approximated negative log-likelihood of JEVAE is still on average lower than the conventional VAE. As most current state-of-the-art VAEs [30] only support regular image datasets, for fair comparison we only compare against known VAEs that are tuned towards spatiotemporal data. It is noted that there is high standard deviation associated with the negative log-likelihood scores due to presence of outlier trials and outlier subjects. Subjects with significant deviation across trials from the overall learnt distribution would face larger scores, further revealing the presence of intra-subject variations in the EEG paradigm which may negatively affect data quality and subsequent classifier training.

In Table I, for the EEG data, the results shows that the JEVAE consistently outperforms the VAE counterparts with 16 features, significantly improving the performance reaching $\mathcal{L} = -74.8$. This is significantly higher compared to the best VAE score of $\mathcal{L} = -87.2$ with a traditional VAE with 8 latent features. At lower numbers of feature levels, JEVAE performed worse with the exception of the $\beta-$JEVAE. At a lower number of features, the JEVAE might be under-parameterised due to the splitting of the latent features and as a result performs poorer compared to the VAE counterparts. As expected, adding features generally improved the performance of the JEVAE models compared to the vanilla VAE. This suggests that JEVAE may scale better with the adding of features due to the splitting of features between two separate distributions for better representation generalisation. Using $\alpha$ and $\beta$ hyperparameters provide significant performance benefits. The addition of a gradient clip yielded poorer results with the exception of the VAE at 16 features, suggesting that exploding gradients is not an issue when training VAEs in this context and may even be detrimental towards optimal learning of the models.

Although it is generally recognized that the addition of a greater number of features would generally lead to an increase in the VAE reconstruction performance [39], we note that the effects plateau off past 16 features. In addition, overly paramaterizing the VAE network may even become detrimental for the model's performance on the evaluation set while concurrently increasing the overall computational cost [45] of training the network. This may be explained by the effect of intra-subject and inter-subject variations on the dataset. The VAE while attempting to learn a large amount of latent features would overfit towards the features relating to the training set. This is due to the training method of the VAE which encourages reconstruction of the input, causing the network to learn parameters that best reconstruct what it has already seen, losing generalizability to the unseen evaluation set due to inter-subject domain differences. Furthermore, too many latent features will result in categorical redundancy [46], whereby the features learnt are not useful towards classification of the motor imagery signal. This causes many of the latent feature points to contain noisy data which results in poorer-defined latent representations.

We also show that the loss function proposed in this study achieves the best model performance across both datasets as shown in table II achieving $\mathcal{L} = -74.8$ for the EEG dataset. When using the conventional reconstruction loss and the sum of individual network reconstruction losses, the model

TABLE III
COMPARISON OF AVERAGE PERFORMANCE (NEGATIVE LOG-LIKELIHOOD) ACROSS DIFFERENT METHODS ON EEG DATA

| Methodology | Mean (SD) | Median | Range |
|---|---|---|---|
| VAE (2014) | 87.17 (±120.94) | 38.69 | 747.73 (11.15-758.88) |
| Normalising Flow VAE (2015) [47] | 89.19 (±123.09) | 43.50 | 750.11 (8.50-758.61) |
| beta-VAE (2016) [36] | 86.21 (±114.90) | 34.83 | 649.18 (10.15-659.33) |
| Conditional Planar Flow VAE (2020) [48] | 99.01 (±139.00) | 43.93 | 858.49 (11.23-869.72) |
| Planar Flow VAE-LSTM (2022) [49] | 126.22 (±201.88) | 55.96 | 1357.98 (16.64-1374.62) |
| **$\beta$-JEVAE (Ours)** | **74.83 (±91.56)** | **33.94** | **478.96 (10.35-489.91)** |

performed worse in the EEG dataset achieving $\mathcal{L} = -84.2$ and $\mathcal{L} = -101.9$ respectively. The individual sum loss performed worst, suggesting that for data with low signal-to-noise ratio it is not ideal to separate the embedding section losses. Overall, this indicates that we do not face performance loss when optimising solely the second embedding section even for high signal-to-noise ratio data. We can therefore strongly infer that the best optimisation strategy would be to optimise the reconstruction ability of solely the second network embedding section as shown in the earlier equation 10. Finally, we compare the proposed method against other common and state-of-the-art methods used to perform representation learning on time-series related data (Table III).

Common methods to train VAEs to learn representations of time-series data include the use of flow-based VAEs [47], [48], [49]. Typically, most flow-based VAEs leverage on the use of normalizing flows [47] in order to perform a density approximation of $p(x)$. This is achieved via the transformation of the simple distribution obtained in a normal VAE into a complex distribution through the implementation of a series of invertible transformation functions on the distribution [50]. In the work proposed, the simple distribution is similarly transformed into a complex distribution to better reflect the true distribution of the input signals. However, rather than applying functions to transform a single distribution to a different complex distribution, the proposed method aims to approximate the true complex distribution via estimated decomposition of the original distribution into that of simple distributions that best reflect the true distribution. Low median and range scores in comparison against flow-based methods (Table III) further display the methods ability to generalize across different subjects despite physiological differences.

The main benefit behind this approach is that the features learnt by the model can be separated into different distributions rather than constrained into a single complex distribution. Additionally, normalizing flows tend to focus on low-level features compared to generalizing towards the broader semantic contents of the input [51], which leads to poorer detection of anomalies and worse generalization towards new target trials with different latent domains. By keeping the decomposed distributions in JEVAE as simple distributions, the decomposed features learnt are able to improve generalization towards learning overall features that capture the underlying signal domains (Table III), which in this case would be defined by the offline and online learning phases, as well as the motor imagery classes of left against right (Fig. 8).

### B. Feature-Informed Transfer Learning

Given the insights obtained from the extracted features, we showcase the effectiveness of the proposed JEVAE by utilizing the features to perform feature-informed transfer learning (Table IV). We implement feature-informed transfer learning via utilizing JEVAE as follows. The JEVAE first learn model parameters that extract the features of the known trials as well as the portion of data to be used for transfer learning. Following which, the first 10 evaluation trials are utilized as an input query to compute the negative log-likelihood of JEVAE via equation 10. A threshold is then used to determine the maximum allowable distance between the evaluation trials against the transfer learning set. Subjects below the allowable threshold would therefore be hypothesized to benefit more from transfer learning and fine-tuning will be executed, while those above the threshold are excluded from fine-tuning. The evaluation of the resultant classifier with or without transfer learning is then determined via the remaining 90 held-out trials from the test set.

From the results (Table IV), the proposed methodology when used in conjunction with transfer learning is able to achieve state-of-the-art performance over other methods which uses adaptive models [1], [17], [53], [56], [58], [59], with significant improvements to the mean accuracy across subjects and an overall higher median. The minimum classification accuracy of the fine-tuned classifier is raised significantly as well. This strongly indicates the usefulness of the proposed JEVAE in minimizing the effect of negative post-adaptation outcomes, which has not been addressed by previous methods on motor imagery classification (Table IV).

## VI. FOUR-CLASS MOTOR IMAGERY

We further validate the effectiveness of the proposed JEVAE on the BCI Competition IV 2a dataset [60]. The dataset consists of EEG recordings from nine healthy participants, including five males and four females. The participants were instructed to perform four different motor imagery tasks as follows, left hand (class 1), right hand (class 2), both feet (class 3), and tongue (class 4). The EEG signals were recorded using a 22-channel EEG amplifier. Two sessions on different days were recorded for each subject. Each participant performed 288 trials across the four classes for each session. The EEG signals were recorded at a sampling rate of 250 Hz, and each trial lasted for 4 seconds. The signals were bandpass filtered between 0.5 and 100 Hz and notch filtered at 50 Hz. For this study, we consider the left and right motor imagery

TABLE IV
COMPARISON BETWEEN AVERAGE CLASSIFICATION ACCURACY (%) AND STANDARD DEVIATION OF DIFFERENT METHODS. THE PROPOSED METHODOLOGY ACHIEVES STATE-OF-THE-ART WITHOUT THE NEED FOR TARGET SUBJECT ADAPTATION AS COMPARED TO PREVIOUS METHODS

| Methodology | Mean (SD) | Median | Range (Max-Min) |
|---|---|---|---|
| MIN2Net (2022) [52] | 72.03 (±14.04) | 72.00 | 55.50 (100.00-44.50) |
| Mutual Inference (2021) [53] | 73.32 (±13.55) | 74.00 | 51.00 (98.00-47.00) |
| Spectral-Spatial CNN (2019) [54] | 74.15 (±15.83) | 75.00 | 60.00 (100.00-40.00) |
| TSMNet (2022) [55] | 74.60 (±14.22) | 73.00 | 53.00 (99.00-46.00) |
| EEG-GAN (2018) [56] | 81.03 (±9.97) | 81.85 | 41.73 (98.77-57.04) |
| DCGAN (2020) [57] | 81.85 (±9.85) | 81.40 | 39.84 (98.52-58.68) |
| Cycle-GAN (2022) [58] | 82.34 (±8.98) | 81.52 | 34.05 (99.08-65.03) |
| Deep CNN (2017) [34] | 84.19 (±9.98) | 84.50 | 47.50 (99.50-52.00) |
| EEGSym (2022) [59] | 84.72 (±11.73) | 82.50 | 40.00 (100.00-60.00) |
| Deep Subject-Adaptive CNN (2021) [1] | 86.89 (±11.41) | 88.50 | 44.00 (100.00-56.00) |
| **JEVAE Feature-Informed Transfer Learning (Ours)** | **87.54 (±10.21)** | **89.00** | **37.00 (100.00-63.00)** |

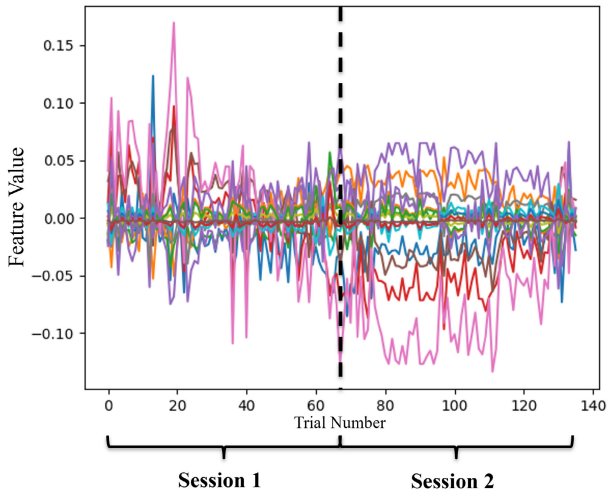


Fig. 4. Learned latent features from a specific subject. Differing trends in the features between the first and second sessions indicate strong inter-session variability and potentially poor adaptation results. Each color represents a single feature variable.

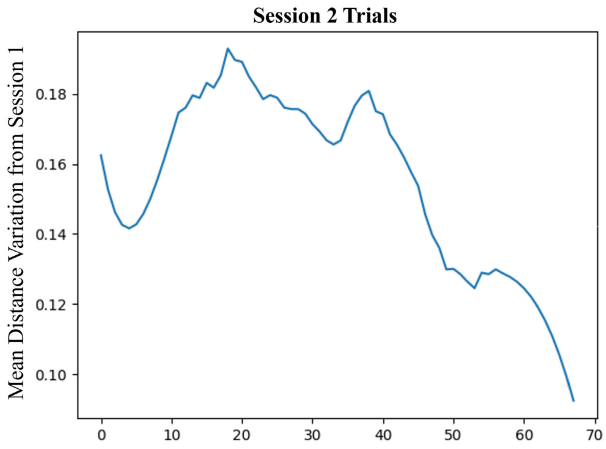


Fig. 5. Savitzsky-Golay moving average of the euclidean distance between session 2 trials against the trials in session 1. A higher distance indicates further deviation from session 1.

trials across both sessions, removing invalid trials for each subject.

As with the Korea University dataset [41], this dataset consists of separate EEG signal collection sessions, making it an appropriate dataset for further validation of the proposed JEVAE. As such, it is expected that there will be instances in subjects that display higherly inter-session signal variability as a result of time variation [8]. We further showcase using the dataset [60] that even between subjects, the presence of inter-session variability is highly subject dependent and therefore difficult for current conventional neural network training paradigms to consider and circumvent this complexity.

The distribution of the learnt features across trials display the proposed method's capability in identifying subjects with higher time-to-time variations with respect to their EEG signals (Fig. 4). As is expected, certain subjects faced larger variations when comparing the EEG signal features between the two separate sessions which may be attributed to the non-stationary shifts in the individual subject's neurophysiological state [61] which may arise due to neuromodulation [62]. While within each of the sessions, we show that the features learnt remained relatively stable since we do not expect to see a significant shift in the subject's state within the same recording sitting, lending further credibility to the efficacy of the proposed method. Observing the Savitzsky-Golay moving average euclidean distance of the second session trials as compared to the first session, the trials show a significant deviation across majority of the trials while largely maintaining consistency within the sessions (Fig. 5).

On the other hand, subjects with high stability in EEG signals show less variations in their features across different sessions (Fig. 6. In these subjects, the Savitzsky-Golay moving average show that the variations of majority of session 2 trials follow closely to those in session 2 (Fig. 7), with the exception of the final few trials which the extracted feature values show a large deviation from.

Comparing against previous methods on the dataset (Table V), using the proposed feature-informed domain extraction methodology outperforms previous methods. Earlier methods such as FBCSP [11] and interval-valued aggregate functions [20] rely on traditional algorithms which lack the computational leverage that modern neural networks offers.

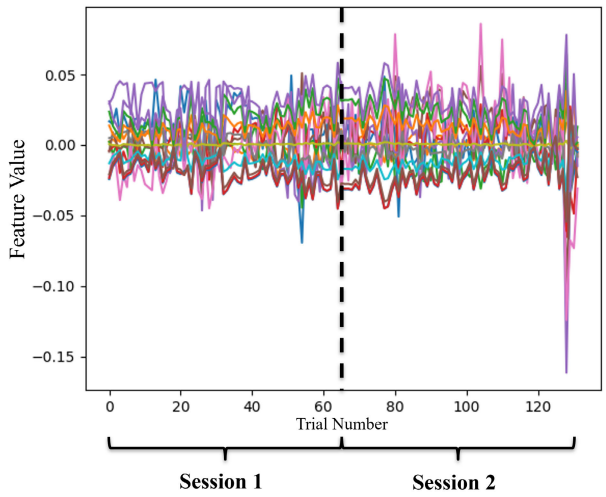


Fig. 6. Learned latent features from a specific subject. Similar trends in the features between the first and second sessions indicate better generalizability between session trials. Each color represents a single feature variable.

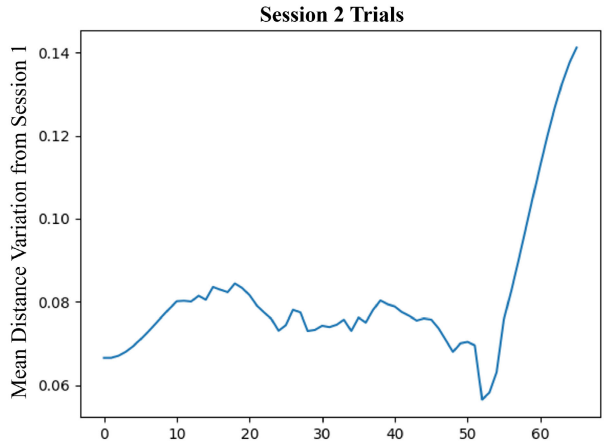


Fig. 7. Savitzsky-Golay moving average of the euclidean distance between session 2 trials against the trials in session 1. A higher distance indicates further deviation from session 1.

TABLE V
COMPARISON BETWEEN AVERAGE CLASSIFICATION ACCURACY (%) ON THE BCI COMPETITION IV 2A DATASET FOR EEG-BASED FOUR-CLASS MOTOR IMAGERY

| Methodology | Accuracy (%) |
|---|---|
| FBCSP (2008) [11] | 66.13 |
| Interval-Valued Aggregate Functions (2021) [20] | 69.43 |
| FBCNet (2021) [12] | 71.53 |
| EEGNet (2018) [63] | 72.32 |
| Tensor-CSPNet (2022) [64] | 73.61 |
| Feature-Informed Domain Extraction (Ours) | 74.42 |

Newer methods based on convolutional neural networks [12], [64] find greater success in integrating the strengths of the previous algorithms into the network architecture. Finally, we show that implementing subject-adaptive ideology by utilizing representation learning can achieve superior results. State-of-the-art performance may be achieved by fine-tuning the network model without the need for any target labels. As such, the network learns spatial-temporal domain features which are specific to the evaluation subject compared to generic features, enabling the classification capabilities of the network on the subject to be increased.

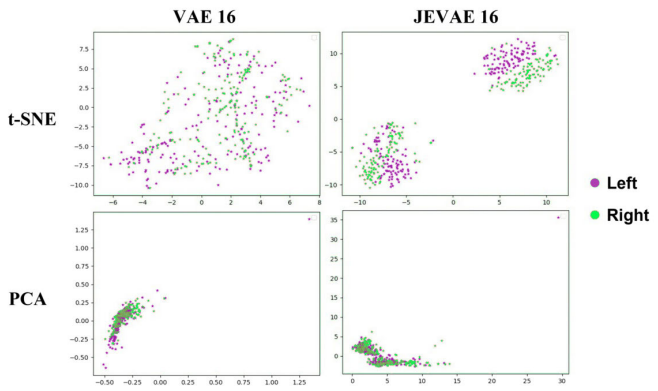


Fig. 8. (a) t-SNE and (b) PCA qualitative comparison between vanilla VAE and Joint Embedding VAE trained on EEG data. The colors indicate the true class label. Separate clusters of the t-SNE and PCA plots by JEVAE display its ability to detect differences in latent domain features of offline and online recordings.

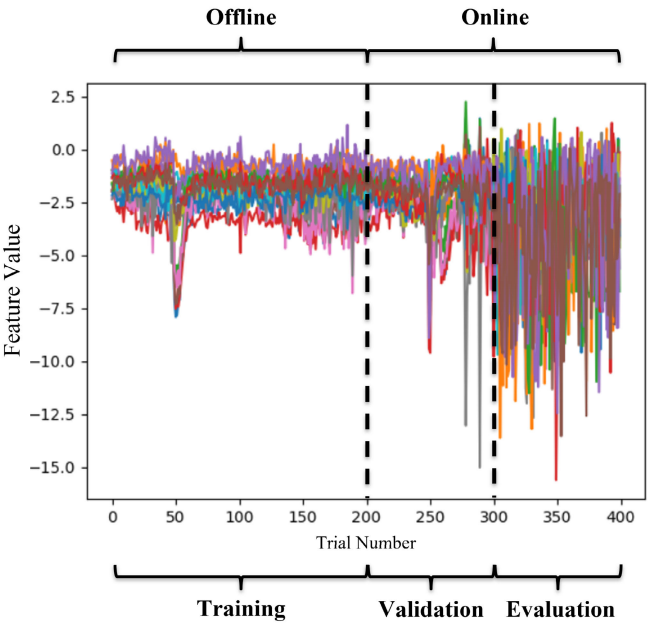


Fig. 9. Learned features across 400 trials. First 200 trials represent the offline session and the last 200 trials represent the online session. Each session consists of two 100 trial phases. Each color represents a single feature variable.

## VII. DISCUSSION

### A. Understanding Poor Adaptation Performance in Subjects

A qualitative study is done by plotting out the principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) [65] plots as shown in Figure 9. The true class labels were used to indicate the left and right motor imagery trials. Since JEVAE splits the features between the two individual network models, to obtain the overall features the two sets of features are concatenated to give the same feature representation size as the traditional VAE counterpart.

Observing the PCA plots in Figure 8(b) between the traditional VAE against the JEVAE, it can be seen that the

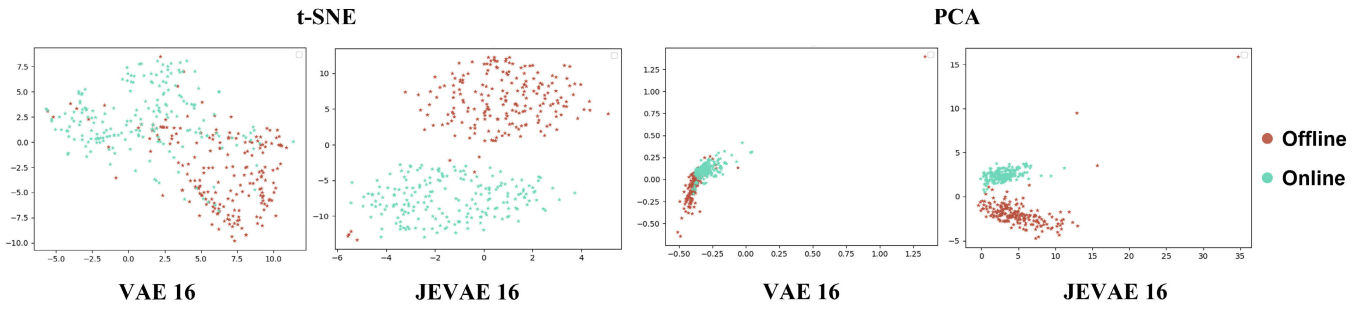


Fig. 10. (a) t-SNE and (b) PCA comparison between vanilla VAE and JEVAE on classifying EEG collection settings. The colors indicate the true class label.

JEVAE offers a clearer split between the features' classes. This is further corroborated by the t-SNE plots in Figure 8(a), whereby the JEVAE is able to show clear separation between the left and right motor imagery features. The JEVAE is even able to capture the features that represent the offline or online domain during the data collection phase. This is shown via the clear distinct separation of two separate clusters with equal number of left and right label classes. Thus, this gives a strong indication that JEVAE is able to learn better distinguishable features among the trials.

### B. Intra-Subject Inter-Session Variation in Key Features

Finally, we study the features of $z_i \sim q(z_i|\cdot)$ learned across 400 trials for a single chosen subject. Figure 9 and 10 shows an observable difference in the general features between each data recording phase. Between each of the recording phases, there are noticeable changes to the general pattern of the learned features for every 100 trials. As expected, the largest observable change can be seen when comparing the features in the offline phase to the online phase. The changes within the sessions are likely to be smaller due to the same recording setup used, however across sessions, users are likely to experience vastly different feedback leading to a large change.

In the context of brain-computer interface (BCI), The JEVAE's ability to extract the features of the EEG signals further corroborate with previous research that the non-stationary features play an important role in defining EEG motor imagery signals [66]. From phase to phase, the features unrelated to motor imagery show clear differentiation. Thus, this study sheds light as to why training a discriminator on different data splits or subjects may result in vastly different results.

Subjects with similar non-stationary features across signal recording phases (Fig. 11) would show better performance compared to subjects with widely varying features [1]. In such cases, the latent domain shift between trials is smaller and thus earlier trials form a closer approximation to the current target trial. Considering that the validation trials come from the same session as the evaluation set, subjects whose variations of latent features in the online session remain small would thus have a higher likelihood of reaching better convergence and generalizability towards the target. Compared to subjects whose features vary greatly within the online session (Fig. 9), the fine-tuning would mistakenly converge towards the latent domain of the validation trial which does not generalize well

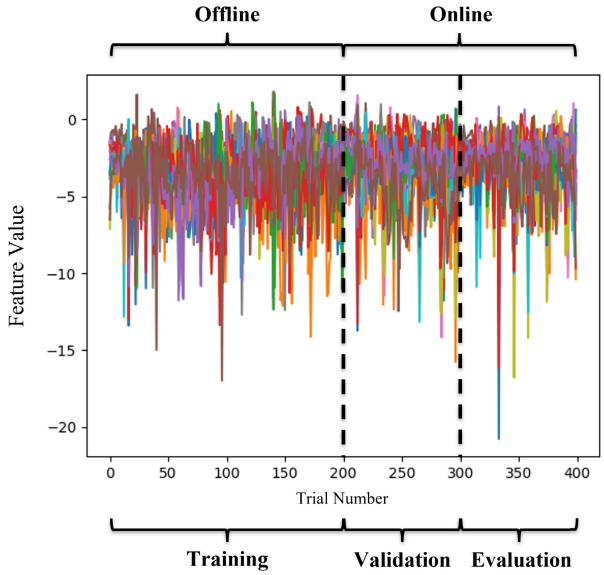


Fig. 11. Learned latent feature representation of a subject. Similar patterns in features across the different EEG recording phases indicate the subject's data is highly homogeneous. The convergence of the fine-tuned model will be closer towards the target.

towards the target domain. Therefore, when training motor imagery discriminators [1], [67], [68], it is vital to take into consideration these features to ensure that the discriminator is trained with a homogeneous domain (Fig. 11) so that it does not get confused by the unrelated features leading to poorer performance.

In previous studies, the distribution of the data across trials for a subject was assumed to be from the same continuous distribution [1], [67], [68]. Based on this assumption, a transfer-learning model that uses earlier trials of the target subject for fine-tuning was proposed [1]. However, the proposed method in this study reveals that this assumption does not hold true across all subjects (Fig. 9) and may even be detrimental towards subject decoding performance. The features extracted using the proposed methodology shows that across offline and online trials, there is a significant difference between the latent features. This strongly indicates that the trials from the offline phase have latent domains that do not coincide with the online phase, which would lead to worse decoding performance if the model is trained upon offline data and evaluated on online data.

Additionally, even adaptation by using the same session data may not necessarily yield improved results. When inspecting

the learnt latent features within each offline and online session (Fig. 9, 4), it can be deduced that there still exists strong intra-session variability. In the case of the online session from Figure 9, the learnt latent features vary greatly from the trials at the start of the session compared to the trials towards the end of the session. This reflects the significance behind the non-stationary variations of feature domains in time series signals such as EEG [69]. When building an adaptive framework for online neural decoding, simply blanket choosing earlier trials for fine-tuning towards later target trials would therefore have negative effects on model performance for subjects displaying such time-to-time variations. Therefore, the feature extraction method proposed in this study serves as an important guide towards detecting these latent domain shifts in EEG signals, enabling models to become more robust against such variability.

JEVAE is therefore a useful tool in discriminating between trials that follow closely to the earlier trials, whether across sessions or even within the session itself. JEVAE enables the specific identification of individual subjects that have high EEG signal variability and therefore may not benefit from adaptation using earlier trials. This ability is useful in common instances whereby neural networks may rely on the assumption that the EEG signal data used for training or fine-tuning are drawn from the same continuous distribution, therefore allowing for the training data of the networks to be curated via feature-informed methodologies such as JEVAE.

### C. Potential Implications

Besides having the ability to detect potentially poor adaptation outcomes, the proposed JEVAE also has an added benefit of detecting anomalous trials. Anomalous trials in EEG motor imagery classification tasks arise when the signals collected in such trials either contain high amounts of signal artefacts or incorrect execution of instructions [70]. In such cases, the EEG signals collected from these trials would form an outlier against the overall distribution of trials. However, due to the abstract nature of biophysical signals, such outliers may be difficult to detect when there is no obvious outward sign. This differs from conventional image datasets such as MNIST [71] or OMNIGLOT [72] where it is a simple task to visually check for any errors. Variational autoencoders such as JEVAE are useful for these types of data as they offer an additional benefit in having the ability to identify anomalous data [73]. Thus, JEVAE through learning the general representation of the signals across all trials would be able to serve as a good indicator of when a trial within the subject stands out from among the rest (Fig. 12).

Beyond motor imagery classification, JEVAE may potentially be applied towards other areas of EEG signal classification tasks such as emotion classification [74], imagined speech intent classification [75], and inner speech decoding [76]. As JEVAE proposes a model framework that encourages better latent feature extraction by additionally focusing on reducing the lossy nature of the general VAE architecture, we hypothesize that JEVAE will perform exceptionally well in tasks that heavily involve reconstructing the original input. For instance in inner speech decoding [76], the network may be trained to reconstruct the original true sentence that the user is tasked to read or imagine. As such, the network's functional loss may be replaced towards the reconstruction loss between the original sentence and the decoded sentence given the EEG signal input. The loss function in such a case would be able to leverage against the architecture of the JEVAE which is highly suitable towards minimizing reconstruction loss without the need for labels.

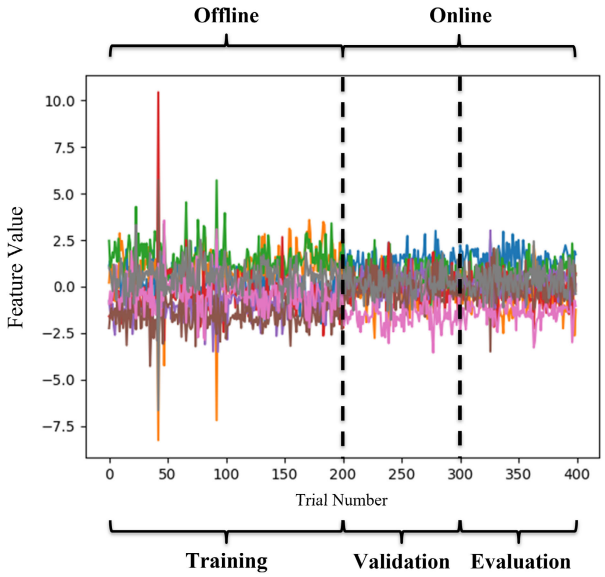


Fig. 12. Learned latent feature representation of a subject. Spikes across the trials indicate the presence of anomalous trial data which does not belong to the same distribution as the overall trials learned.

## VIII. Conclusion

In summary, this study introduces JEVAE, a novel Variational Autoencoder (VAE) framework, designed to enhance lower bounds and improve log-likelihood approximations. JEVAE employs a secondary autoencoder network to optimize reconstruction ability based on a tailored input probability distribution $x'$. The loss function is simplified to optimize the latest embedding section. Additionally, the partitioning of latent variables across two networks grants flexibility in network design. By representing variables in separate spaces, JEVAE enhances the learning of probability distributions in the latent space, presenting a model-agnostic approach that can extend to other variational inference-based networks [77]. This advancement enables state-of-the-art performance and opens new possibilities for feature-informed Brain-Computer Interface (BCI) applications.

However, JEVAE inherits limitations from VAE networks and their variants, lacking a definitive methodology for determining the ideal network structure for each embedding section. Achieving optimal JEVAE performance requires individual optimization of embedding section structures, considering the interplay of each section on the entire algorithm. Hence, architectural optimization often necessitates an empirical approach. Future research could delve into understanding the dependency of the first autoencoder network in JEVAE on the second, laying a foundational framework for creating an overarching network structure adaptable to various paradigms.

## REFERENCES


[1] K. Zhang, N. Robinson, S.-W. Lee, and C. Guan, "Adaptive transfer learning for EEG motor imagery classification with deep convolutional neural network," *Neural Netw.*, vol. 136, pp. 1–10, Apr. 2021.

[2] K. Belwafi, S. Gannouni, H. Aboalsamh, H. Mathkour, and A. Belghith, "A dynamic and self-adaptive classification algorithm for motor imagery EEG signals," *J. Neurosci. Methods*, vol. 327, Nov. 2019, Art. no. 108346.

[3] K. K. Ang and C. Guan, "EEG-based strategies to detect motor imagery for control and rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 4, pp. 392–401, Apr. 2017.

[4] L. Zhu et al., "Multi-source fusion domain adaptation using resting-state knowledge for motor imagery classification tasks," *IEEE Sensors J.*, vol. 21, no. 19, pp. 21772–21781, Oct. 2021.

[5] G. Pfurtscheller, C. Neuper, A. Schlogl, and K. Lugger, "Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters," *IEEE Trans. Rehabil. Eng.*, vol. 6, no. 3, pp. 316–325, Sep. 1998.

[6] N. F. Ince, F. Goksu, A. H. Tewfik, and S. Arica, "Adapting subject specific motor imagery EEG patterns in space-time-frequency for a brain computer interface," *Biomed. Signal Process. Control*, vol. 4, no. 3, pp. 236–246, Jul. 2009.

[7] A. Abu-Rmileh, E. Zakkay, L. Shmuelof, and O. Shriki, "Co-adaptive training improves efficacy of a multi-day EEG-based motor imagery BCI training," *Frontiers Human Neurosci.*, vol. 13, p. 362, Oct. 2019.

[8] S. Saha and M. Baumert, "Intra- and inter-subject variability in EEG-based sensorimotor brain computer interface: A review," *Frontiers Comput. Neurosci.*, vol. 13, p. 87, Jan. 2020.

[9] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," 2015, *arXiv:1511.06448*.

[10] J. Ma, B. Yang, W. Qiu, Y. Li, S. Gao, and X. Xia, "A large EEG dataset for studying cross-session variability in motor imagery brain–computer interface," *Sci. Data*, vol. 9, no. 1, p. 531, Sep. 2022.

[11] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain–computer interface," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 2390–2397.

[12] R. Mane et al., "FBCNet: A multi-view convolutional neural network for brain–computer interface," 2021, *arXiv:2104.01233*.

[13] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[14] M. Noroozi, H. Pirsiavash, and P. Favaro, "Representation learning by learning to count," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5899–5907.

[15] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," 2019, *arXiv:1906.02691*.

[16] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5400–5409.

[17] G. Zoumpourlis and I. Patras, "Motor imagery decoding using ensemble curriculum learning and collaborative training," 2022, *arXiv:2211.11460*.

[18] C. Chen, Z. Chen, B. Jiang, and X. Jin, "Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation," in *Proc. Nat. Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 3296–3303.

[19] H. W. Ng and C. Guan, "Efficient representation learning for inner speech domain generalization," in *Proc. Int. Conf. Comput. Anal. Images Patterns*. Cham, Switzerland: Springer, 2023, pp. 131–141.

[20] J. Fumanal-Idocin et al., "Interval-valued aggregation functions based on moderate deviations applied to motor-imagery-based brain–computer interface," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 7, pp. 2706–2720, Jul. 2022.

[21] C. Feichtenhofer, H. Fan, B. Xiong, R. Girshick, and K. He, "A large-scale study on unsupervised spatiotemporal representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3298–3308.

[22] G. Zhong, L.-N. Wang, X. Ling, and J. Dong, "An overview on data representation learning: From traditional feature learning to recent deep learning," *J. Finance Data Sci.*, vol. 2, no. 4, pp. 265–278, Dec. 2016.

[23] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193907–193934, 2020.

[24] Z. Liu, Y. Lin, and M. Sun, *Representation Learning for Natural Language Processing*. Singapore: Springer, 2020.

[25] X. Lagorce, S.-H. Ieng, X. Clady, M. Pfeiffer, and R. B. Benosman, "Spatiotemporal features for asynchronous event-based data," *Frontiers Neurosci.*, vol. 9, p. 46, Feb. 2015.

[26] F. Perez-Cruz, "Kullback–Leibler divergence estimation of continuous distributions," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 1666–1670.

[27] M. Kim and V. Pavlovic, "Recursive inference for variational autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19632–19641.

[28] E. J. McShane, "Jensen's inequality," *Bull. Amer. Math. Soc.*, vol. 43, no. 8, pp. 521–527, 1937.

[29] L. He, B. Liu, D. Hu, Y. Wen, M. Wan, and J. Long, "Motor imagery EEG signals analysis based on Bayesian network with Gaussian distribution," *Neurocomputing*, vol. 188, pp. 217–224, May 2016.

[30] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[31] A. Klushyn, N. Chen, R. Kurle, B. Cseke, and P. van der Smagt, "Learning hierarchical priors in VAEs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.

[32] A. Vahdat and J. Kautz, "NVAE: A deep hierarchical variational autoencoder," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19667–19679.

[33] J. D. Havtorn, J. Frellsen, S. Hauberg, and L. Maaløe, "Hierarchical VAEs know what they don't know," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4117–4128.

[34] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.

[35] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Trans. Circuits Syst.*, vol. 8, no. 7, pp. 579–588, 2009.

[36] I. Higgins et al., "$\beta$-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. ICLR*, 2016, pp. 1–22.

[37] T. Rainforth et al., "Tighter variational bounds are not necessarily better," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4277–4285.

[38] J. M. Tomczak and M. Welling, "VAE with a VampPrior," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 1214–1223.

[39] Y. Zhou, X. Liang, W. Zhang, L. Zhang, and X. Song, "VAE-based deep SVDD for anomaly detection," *Neurocomputing*, vol. 453, pp. 131–140, Sep. 2021.

[40] F. Ye and A. G. Bors, "Deep mixture generative autoencoders," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5789–5803, Oct. 2022.

[41] M.-H. Lee et al., "EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy," *GigaScience*, vol. 8, no. 5, May 2019, Art. no. giz002.

[42] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain–computer communication," *Proc. IEEE*, vol. 89, no. 7, pp. 1123–1134, Jul. 2001.

[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[45] B. Dai, Y. Wang, J. Aston, G. Hua, and D. Wipf, "Hidden talents of the variational autoencoder," 2017, *arXiv:1706.05148*.

[46] X. Li, Z. Xu, K. Wei, and C. Deng, "Generalized zero-shot learning via disentangled representation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, 2021, pp. 1966–1974.

[47] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1530–1538.

[48] R. Selvan, F. Faye, J. Middleton, and A. Pai, "Uncertainty quantification in medical image segmentation with normalizing flows," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, 2020, pp. 80–90.

[49] X.-B. Jin, W.-T. Gong, J.-L. Kong, Y.-T. Bai, and T.-L. Su, "PFVAE: A planar flow-based variational auto-encoder prediction model for time series data," *Mathematics*, vol. 10, no. 4, p. 610, Feb. 2022.

[50] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 2617–2680, 2021.

[51] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Why normalizing flows fail to detect out-of-distribution data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 20578–20589.

[52] P. Autthasan et al., "MIN2Net: End-to-end multi-task learning for subject-independent motor imagery EEG classification," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 6, pp. 2105–2118, Jun. 2022.

[53] E. Jeon, W. Ko, J. S. Yoon, and H.-I. Suk, "Mutual information-driven subject-invariant and class-relevant deep representation learning in BCI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 2, pp. 739–749, Feb. 2023.

[54] O.-Y. Kwon, M.-H. Lee, C. Guan, and S.-W. Lee, "Subject-independent brain–computer interfaces based on deep convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 3839–3852, Oct. 2020.

[55] R. J. Kobler, J.-I. Hirayama, Q. Zhao, and M. Kawanabe, "SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG," 2022, *arXiv:2206.01323*.

[56] K. G. Hartmann, R. T. Schirrmeister, and T. Ball, "EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals," 2018, *arXiv:1806.01875*.

[57] F. Fahimi, S. Dosen, K. K. Ang, N. Mrachacz-Kersting, and C. Guan, "Generative adversarial networks-based data augmentation for brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 4039–4051, Sep. 2021.

[58] K. Yin, B.-H. Lee, B.-H. Kwon, and J.-H. Cho, "Target-centered subject transfer framework for EEG data augmentation," 2022, *arXiv:2212.00723*.

[59] S. Pérez-Velasco, E. Santamaría-Vázquez, V. Martínez-Cagigal, D. Marcos-Martínez, and R. Hornero, "EEGSym: Overcoming inter-subject variability in motor imagery based BCIs with deep learning," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1766–1775, 2022.

[60] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI competition 2008—Graz data set A," *Inst. Knowl. Discovery, Lab. Brain Comput. Interfaces, Graz Univ. Technol.*, vol. 16, pp. 1–6, Jan. 2008.

[61] S. Saha, K. I. U. Ahmed, R. Mostafa, L. Hadjileontiadis, and A. Khandoker, "Evidence of variabilities in EEG dynamics during motor imagery-based multiclass brain–computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 371–382, Feb. 2018.

[62] U. Ziemann and H. R. Siebner, "Inter-subject and inter-session variability of plasticity induction by non-invasive brain stimulation: Boon or bane?" *Brain Stimulation*, vol. 8, no. 3, pp. 662–663, May 2015.

[63] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.

[64] C. Ju and C. Guan, "Tensor-CSPNet: A novel geometric deep learning framework for motor imagery classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10955–10969, Dec. 2023.

[65] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[66] X. Yu, P. Chum, and K.-B. Sim, "Analysis the effect of PCA for feature reduction in non-stationary EEG based motor imagery of BCI system," *Optik*, vol. 125, no. 3, pp. 1498–1502, Feb. 2014.

[67] X. Hong et al., "Dynamic joint domain adaptation network for motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 556–565, 2021.

[68] J. Han, X. Gu, and B. Lo, "Semi-supervised contrastive learning for generalizable motor imagery EEG classification," in *Proc. IEEE 17th Int. Conf. Wearable Implant. Body Sensor Netw. (BSN)*, Jul. 2021, pp. 1–4.

[69] J. Faller, C. Vidaurre, T. Solis-Escalante, C. Neuper, and R. Scherer, "Autocalibration and recurrent adaptation: Towards a plug and play online ERD-BCI," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 20, no. 3, pp. 313–319, May 2012.

[70] A. K. Porbadnigk et al., "When brain and behavior disagree: Tackling systematic label noise in EEG data with machine learning," in *Proc. Int. Winter Workshop Brain Comput. Interface (BCI)*, Feb. 2014, pp. 1–4.

[71] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.

[72] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, Dec. 2015.

[73] J. Xu, Y. Zheng, Y. Mao, R. Wang, and W.-S. Zheng, "Anomaly detection on electroencephalography with self-supervised learning," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 363–368.

[74] M. Li and B.-L. Lu, "Emotion classification based on gamma-band EEG," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sep. 2009, pp. 1223–1226.

[75] K. Brigham and B. V. K. V. Kumar, "Imagined speech classification with EEG signals for silent communication: A preliminary investigation into synthetic telepathy," in *Proc. 4th Int. Conf. Bioinf. Biomed. Eng.*, Jun. 2010, pp. 1–4.

[76] N. Nieto, V. Peterson, H. L. Rufiner, J. E. Kamienkowski, and R. Spies, "Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition," *Sci. Data*, vol. 9, no. 1, p. 52, Feb. 2022.

[77] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," 2015, *arXiv:1509.00519*.