IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 September 2022, accepted 29 September 2022, date of publication 6 October 2022, date of current version 13 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3212777

RESEARCH ARTICLE

Examining the Size of the Latent Space of Convolutional Variational Autoencoders Trained With Spectral Topographic Maps of EEG Frequency Bands

TAUFIQUE AHMED AND LUCA LONGO

Artificial Intelligence and Cognitive Load Research Laboratory, School of Computer Science, Technological University Dublin, Dublin 7, D07 H6K8 Ireland

Corresponding authors: Taufique Ahmed (taufique.ahmed@tudublin.ie) and Luca Longo (luca.longo@tudublin.ie)

This work was supported by the Technological University Dublin, Ireland, under Grant PB04433.

ABSTRACT Dimensionality reduction and the automatic learning of key features from electroencephalographic (EEG) signals have always been challenging tasks. Variational autoencoders (VAEs) have been used for EEG data generation and augmentation, denoising, and automatic feature extraction. However, investigations of the optimal shape of their latent space have been neglected. This research tried to understand the minimal size of the latent space of convolutional VAEs, trained with spectral topographic EEG head maps of different frequency bands, that leads to the maximum reconstruction capacity of the input and maximum utility for classification tasks. Head maps are generated employing a sliding window technique with a 125ms shift. Person-specific convolutional VAEs are trained to learn latent spaces of varying dimensions while a dense neural network is trained to investigate their utility on a classification task. The empirical results suggest that when VAEs are deployed on spectral topographic maps with shape 32×32, deployed for 32 electrodes from 2 seconds cerebral activity, they were capable of reducing the input up to almost 99%, with a latent space of 28 means and standard deviations. This did not compromise the salient information, as confirmed by a structural similarity index, and mean squared error between the input and reconstructed maps. Additionally, along the 28 means maximized the utility of latent spaces in the classification task, with an average 0.93% accuracy. This study contributes to the body of knowledge by offering a pipeline for effective dimensionality reduction of EEG data by employing convolutional variational autoencoders.

INDEX TERMS Electroencephalography, convolutional variational autoencoder, latent space, deep learning, frequency bands, spectral topographic maps, and neural networks.

I. INTRODUCTION

Electroencephalography (EEG) is a technique of recording brain electrical potentials using electrodes placed on the scalp [1]. It is well known that EEG signals contain essential information in the frequency, temporal and spatial domains. For example, some studies have converted EEG signals into topographic power head maps to preserve spatial information [2]. Others have produced spectral topographic head maps of different EEG bands to both preserve information in the spatial domain and take advantage of the information in the frequency domain [3]. However, topographic maps contain highly interpolated data in between electrode locations and are often redundant. For this reason, convolutional neural networks are often used to reduce their dimensionality and learn relevant features automatically [4]. Also, most of these networks are part of larger architectures for classification purposes [5]. However, they often neglect the size of these architectures, as their main goal is to maximize accuracy. This lead often to significant computational time, and redundant training operations. Therefore, we argue that most of these networks can be significantly reduced, trained faster, and lead

The associate editor coordinating the review of this manuscript and approving it for publication was Ludovico Minati.

to models still with high classification accuracy. An Autoencoder (AE) is a deep learning neural network architecture used to learn efficient codings in an unsupervised fashion, without the use of labeled data. These codings, often referred to as latent spaces, are usually of a lower dimension than the original input and they are used to reconstruct it with high fidelity [6]. A Variational Autoencoder (VAE) is a specific type of autoencoder that builds a probabilistic model of the input sample and then reconstructs it using that model. In a VAE, a probabilistic distribution is built to estimate the distribution of the input data. Therefore, it is not a single computation but rather an estimation of the sample data. As a result, VAE can be employed to generate synthetic data [7]. The encoding method of a VAE is identical to that of regular autoencoders. However, the distinction is that VAE transforms the input data into probability distribution parameters, such as a Gaussian's mean and variance. This method creates a continuous, structured latent space that is helpful for the reconstruction of the data as shown in the figure 1.

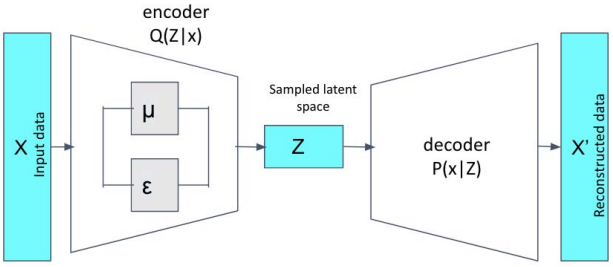

FIGURE 1. The structure of a Variational Autoencoder (VAE) that maps the input data into the parameters of a probability distribution, such as the mean and the variance of a Gaussian distribution.

VAEs have shown a wide application with electroencephalographic (EEG) signals [8], [9], [10]. It has been also employed in various studies for applications in finance, speech/audio source separation, and bio-signals [11]. In these studies, the goal was to learn optimal encoding (latent spaces) that could be used for data generation and augmentation, data denoising, and automatic feature extraction. However, research into VAE devoted to extracting and analysing latent spaces of different dimensions is limited. In one of the studies, researchers investigated a number of latent space dimensions of VAE for generating more relevant EEG features from raw EEG data, improving speech recognition [12]. The VAE-based modelling method outperformed Principal Component Analysis (PCA) in terms of dimensionality reduction. However, its reconstruction performance did not always improve as the number of latent space dimensions was increased [13]. From our literature review, it seems that limited work exists on understanding the maximum dimensionality reduction that can be performed and that can still lead to the preservation of the relevant features and meaning in the EEG data without losing important information. Consequently, it is important to create models that can perform dimensionality reduction in an effective way, without compromising the relationships and semantics within the data. The autoencoder uses convolutional operations over input topographic maps to learn salient high-level features that are lower in dimension and therefore more portable since they require a significant amount of digital memory to be stored. Additionally, this lower dimension contains the relevant and salient representations of EEG data that can be used for various purposes. For example, these include the generation of synthetic EEG topographic head maps for data augmentation, or their use for solving various classification tasks. In this study, the goal is to tackle this research problem and, to determine the minimal size of the latent space of a convolutional VAE, trained with spectral topographic EEG maps of different bands, that leads to maximum reconstruction capacity of the input data, and also maximum utility in classification tasks.

Therefore, the research question being addressed is: *What is the minimum latent space dimension that can be learnt with a convolutional variational autoencoder trained with spectral topographic maps of different EEG bands, that lead to maximum input reconstruction capacity and maximum utility for classification tasks?*

The remainder of the work is structured as follows. Section II investigates related work on VAE used with EEG signals, whereas Section III describes an empirical study and its methodology to answer the above research question. Section IV presents the experimental results and findings. Finally, Section V concludes the manuscript by describing the contribution to the body of knowledge and highlighting future work directions.

II. RELATED WORK

The objective of traditional Autoencoders (AE) is to learn salient latent representations from unlabeled data and ignore irrelevant features. As a result, the reconstructed data will be identical to the input data. The reconstruction procedure is divided into two steps. During the encoding stage, a neural network uses a set of encoding parameters $\theta = \{W, b\}$ to translate the input x to a hidden representation $y = f_\theta(x) = s(Wx + b)$. Secondly, by using decoding parameters $\theta' = \{W', b'\}$, the hidden representation y is mapped to the reconstructed vector $z = g_{\theta'}(y) = s\left(W'y + b'\right)$ [14]. Variational Autoencoders (VAEs) were recently proposed as an effective extension of AEs, for modeling a data's probability distribution and learning a latent space, usually of a lower dimension, always in the absence of explicit supervision [15]. In detail, this latent space is not composed of a fixed vector, but of a mixture of distributions. A VAE allows us to encode an input x to a latent vector $z = Encoder(x) \sim q(z \mid x)$ using an encoder network, and then use another network to decode this latent vector z back to a shape that is as close as possible to the original input data $\bar{x} =$ Decoder $(z) \sim p(x \mid z)$. In other words, the goal is to maximize the marginal log-likelihood of each observation in x, and the VAE reconstruction loss $\mathcal{L}_{rec}$ is the negative anticipated log-likelihood of the observations in x [15] as in the following:

$$\mathcal{L}_{rec} = -\mathbb{E}_{q(z|x)}[\log p(x \mid z)] \quad (1)$$

The presence of hidden components that are integrated to produce visible data is critical to the majority of commonly used methods for inferring latent variables [16]. VAE-based latent space analysis and decoding of EEG signals are important since it can precisely define and determine the latent relevant features [13]. The following sub-section examines previous research on VAE used with EEG signals.

A. VARIATIONAL AUTOENCODERS FOR DATA AUGMENTATION

Classification learning algorithms require adequate samples to successfully build accurate models. However, as in the case of EEG signals, unfortunately, it is not always possible to collect a large amounts of data. To overcome this problem, a VAE can be used as a data generator capable of producing synthetic data that resembles the input samples. For example, researchers demonstrated how a VAE-based approach could be used as a generative model to increase the size of EEG datasets [8]. VAE has been used as an effective computational technique for generating EEG data while a limited amount of participants were exposed to painful stimuli [17]. VAE was also trained as a generative model for improving EEG-based emotion recognition. In detail, the experimental results on two emotion datasets (SEED and DEAP) indicate that with the newly generated synthetic data, classification performance was improved respectively by 10.2% and 5.4% [18]. The Conditional Variational Autoencoder (CVAE) is a VAE extension in which both latent variables and data are conditioned on some random variables. The enriched training datasets created by the CVAE-GAN approach, which combines a CVAE with a GAN for learning latent representations from EEG brain signals, it can greatly enhance the performance of Motor Imagery (MI) EEG recognition [19]. CVAE is also used to generate time-series multichannel signals with spectro-temporal EEG patterns that are expected to be observed under different MI circumstances [20].

B. VARIATIONAL AUTOENCODERS FOR FEATURE REPRESENTATION

In general, the dimensions of the latent space of autoencoders are usually lower than the dimensions of the original data, just as they are for VAE. Furthermore, if the latent space has multiple dimensions, the VAE expression ability and its feature representation capacity will vary. VAE often employs Kullback-Leibler (KL) divergence, which is a measure of how the probability distribution of the latent space differs from that generated by sampling data from it [21]. A special version of the VAE was proposed in [22], focused on learning a generalised model of emotion by concurrently optimizing the goal or learning normally distributed and subject-independent feature representations, via the use of spectral topography data. The ultimate objective was to maximize dataset inter-compatibility, improve robustness to localized electrode noise, and provide a more generally applicable method within neuroscience. Similarly, transfer learning was introduced in brain-computer interface (BCI) research to learn subject-invariant representations by simultaneously training a convolutional VAE and an adversarial network with EEG data collected during Motor Imagery (MI) [23]. Another study has employed VAE for extracting relevant latent representations that are noise-free [24]. Authors have compared their novel method against the FastICA method for noise reduction, and have shown that it can reduce the computational complexity in further EEG processing, besides removing data multicollinearity, which turns out to be beneficial in the subsequent use of such representations, for example in building classification models. Event-Related Potential (ERP) based Driver-Vehicle Interfaces (DVIs) have been created to provide a communication channel for individuals with disabilities to operate a car. However, it was a time-consuming and complex training approach to construct the decoding model that can transform EEG signals into commands. To address this issue, the VAE learning technique was used as a robust feature representation of EEG signals along with a Prior Information-based Transductive Support Vector Machine (PI-TSVM) classifier to translate these features into commands [25].

Another type of VAE is the Introspective Variational Autoencoder (IVAE), which is used for creating high-resolution photographic images. It is capable of self-evaluating the quality of its generated samples and improving itself as a result. It can be used to extract multi-scale features from EEG spectrograms by replacing the main structure in the encoder and decoder layers with an optimal local sparse structure in a convolutional vision network [26]. This approach was used in the context of disease identification. For example, researchers employed a dataset with several disease categories and compared VAE and IVAE with varied latent space dimensions. Authors discovered that the latent representations learned by training an IVAE outperformed VAE in terms of image reconstruction capacity [27]. Always in the context of disease identification, another application focused on extracting and evaluating latent representations with varying latent dimensions from samples such as EEG signals and medical images [10]. Furthermore, studies illustrate how to reduce dimensionality and learn relevant representations using VAE on EEG signals to anticipate cognitive control in older adults [28]. Researchers collected EEG data from many participants while they watched various images appearing on a screen. A deep learning approach was used by employing a Long Short-Term Memory (LSTM) stacked with a VAE devoted to learning a more compact and noise-free representation of EEG data. A model was trained and tested by using EEG data from six subjects while they observed different images from 40 imageNet classes. The performance of the latent space of such a model was evaluated and results indicate that EEG signals contain patterns related to visual content, which can be effectively used to generate images that are semantically coherent with the evoked visual stimuli [29].

C. VARIATIONAL AUTOENCODERS FOR CLASSIFICATION

VAE plays a vital role in automatic emotion recognition and anomaly detection by modeling human observable behaviors such as brain signals, speech, facial expressions, and linguistic content. There has been significant research on multichannel EEG for emotional arousal and speech recognition with unsupervised feature learning using VAE and other generative models [12], [30]. In one work, researchers used raw EEG signals to train a model with a convolutional variational autoencoder in a supervised fashion to predict epileptic seizures in interictal and preictal brain states. To differentiate between the preictal and interictal signals, the latent representation vector Z is fully linked to a single neuron with a sigmoid activation function. For two reasons, the suggested system chose to build the latent vector z with only two dimensions. The first is because the primary goal of the proposed system is to obtain the highest possible classification accuracy even if retraining of the encoder network is required. The second reason is to make it easy to present the system's classification findings through a two-dimensional latent space visualisation [31], [32]. Person authentication based on EEG has become an important tool in modern biometrics. As a result, VAE is utilized to learn the latent representation of EEG signals from users for authentication purposes [33]. One of the recent work focuses on how VAE is used to determine the states of latent variables from multichannel EEG signals that effectively contributes to emotional processing in order to build an emotion recognition model on two public datasets (DEAP and SEED). This model is built to examine the performance of the learned latent space from VAE, AE, and ICA. According to the results of the model, the VAE's latent space outperformed that of AE and ICA, providing more meaningful and effective information for emotional state inference [34].

Despite the wide application and research towards the extraction of relevant information from multi-channel EEG signals and data generalization across participants, the identification of the minimal size latent space dimension utilizing EEG data for reconstruction purposes and secondary utility for classification tasks remains inadequate.

III. RESEARCH DESIGN AND METHODOLOGY

In this study, we anticipate that there exists a specific minimal size of the latent space, learn with a VAE, that is effective both for maximal reconstruction capacity and for maximum accuracy when employed in classification tasks. The detailed design of this research is illustrated in figure 2, and the following sections describe its phases and components.

A. DATASET

1) DEAP

The DEAP dataset has been selected because it contains multi-channel EEG recordings with a good amount of participants and tasks. In fact, EEG data were recorded from 32 people who watched 40 one-minute excerpts of music videos [35]. Each participant was asked to rate a video after watching a 60-second music clip. Each video was graded on a 1–9 scale for dominance, like/dislike, valence, familiarity, and arousal. The standard 10-20 systems was employed with the following 30 electrode positions: 'Fp1', 'AF3', 'F7', 'F3', 'FC1', 'FC5', 'T7', 'C3', 'CP1', 'CP5', 'P7', 'P3', 'Pz', 'PO3', 'O1', 'Oz', 'O2', 'PO4', 'P4', 'P8', 'CP6', 'CP2', 'C4', 'T8', 'FC6', 'FC2', 'F4', 'F8', 'AF4', 'Fp2', 'Fz', 'Cz'. Pre-processing included signal re-sampling at 128 Hz, and a band-pass frequency filter to operate in the 1–45hz frequencies. EOG artifacts were eliminated using a blind source separation technique, as described in [35].

B. EEG DATA PRE-PROCESSING AND TOPOGRAPHIC HEAD MAPS GENERATION

Multichannel EEG data was split into discrete time-slices using a sliding window technique with a 125 ms shift. The first three seconds of the pre-trial baseline are deleted because people are not presented with any video and we are not interested in evaluating their neural responses at this rest state. We employ the fast-Fourier transformation (FFT) on each time-slices to extract the information in the power spectrum. Each power spectrum is divided into the five EEG bands with Delta (0.5-4Hz), Theta (4-8Hz), Alpha (8-12Hz), Beta (12-30Hz), and Gamma (30-45Hz) [36]. Subsequently, the next step is to generate multichannel spatially-preserving topographic EEG head maps from each time-slice for each participant and each video (2. The length of each time-slice was set to different values, respectively 0.5, 1, 1.5, and 2 seconds. The rationale was to experimentally identify the most promising one for optimal latent space formation. Subsequently, the centroid of the frequency amplitude for each band is computed, for a time slice, and all centroids are positioned in a 3D space to produce a scattered head map, one for each band. Additionally, polar projection is applied to each scattered map to produce 2D head maps. Each 2D map is interpolated into five 2D maps, one for each EEG band, subsequently aggregated into a tensor for further processing. Figure 3 depicts these five spatially-preserving topographic maps projected into a 2D map for each EEG band, forming a $32 \times 32 \times 5$ tensor.

C. A CONVOLUTIONAL VARIATIONAL AUTOENCODER

After forming the topographic maps, a Convolutional Variational Autoencoder (CNN-VAE) is constructed, aimed at turning input data into probability distribution parameters, including the mean and standard deviation of a Gaussian distribution. This method provides a continuous, structured latent space that can be subsequently analysed and used for classification purposes. The CNN-VAE design is made up of the following components:

- The encoder is a neural network that takes a tensor of $32 \times 32 \times 5$ dimension (as in figure 3) and defines the approximate posterior distribution $Q(Z \mid x)$, where x is the input tensor and Z is the latent space. Simply by

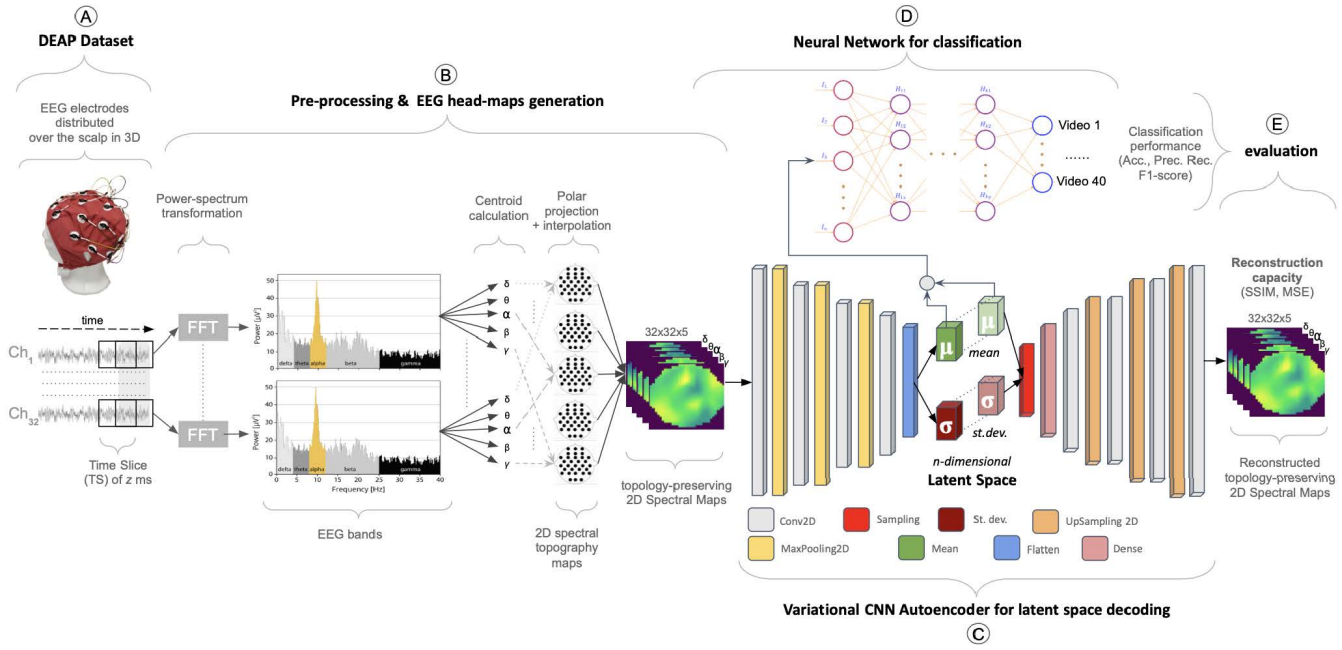

FIGURE 2. A pipeline for EEG data pre-processing and representational learning with convolutional variational autoencoder (CNN-VAE). A)The DEAP dataset was used to build a CNN-VAE from EEG signals. B) EEG signals are segmented into windows. For each signal in a window, FFT is applied to obtain information in the power spectrum for each band (delta, theta, alpha, beta, gamma), and the centroid of the frequency amplitudes is computed for each of them, which produces a scattered head map, one for each EEG band. This scattered spatial-preserving map is then interpolated to produce a full topographic map C) a CNN-VAE model is learned for latent space decoding D) a neural network is used to assess the learned latent space from CNN-VAE E) reconstruction capacity is used as an evaluation metric to assess the VAE, and classification performance to assess its utility to a classification task.

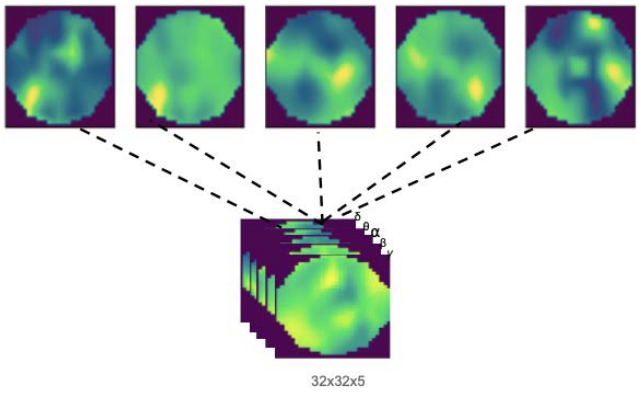

FIGURE 3. Generated spatially-preserving topographic maps projected into a 2D map for each EEG band, forming a 32 × 32 × 5 tensor.

expressing the distribution as a diagonal Gaussian, the network will generate the mean and standard deviation parameters of a factorized Gaussian. The architecture (figure 2, C) consists of four 2D convolutional layers, each of these followed by a max pooling layer to reduce the dimensions of the feature maps. ReLU is used as the activation function in each convolutional layer.

- The decoder of the CNN-VAE is a generative network that takes a latent space Z as input and outputs the parameters for the conditional distribution $P(x \mid Z)$ of the observation (as shown in the right part of figure 2, C). Similarly, like the encoder network, the decoder consists of four 2D convolutional layers, each of these layers followed by an up-sampling layer in order to reconstruct the data up to the shape of the original input. ReLU is used as an activation function in each convolutional layer to regularize the neural network.
- the reparameterization trick is used to generate a sample for the decoder by sampling from the latent distribution defined by the encoder's parameters. As the backpropagation algorithm can not flow through a random sample node in CNN-VAE, sampling operations create a bottleneck. To address this, the reparameterization technique is used to approximate the latent space Z using the decoder parameters along with an additional one, the ϵ parameter [37]:

$$Z = \mu + \sigma \odot \epsilon \tag{2}$$

where μ and σ denote the mean and standard deviation of a Gaussian distribution, whereas the ϵ can be thought of as random noise that is used to keep the stochasticity of Z. The latent space is now generated by a function of μ, σ, and ϵ, allowing the model to backpropagate gradients in the encoder via μ and σ, while maintaining stochasticity via ϵ.

- the CNN-VAEs is optimized via a loss function in order to verify that the latent space is both continuous and complete. This loss function was set as the binary cross-entropy loss paired with the Kullback–Leibler divergence loss, which is a measure of how two probability distributions differ from one

another [37], [38].

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right) \tag{3}$$

This CNN-VAE architecture is trained using data from random 28 videos from a single participant, random 6 for validation, and the remaining 6 videos for testing respectively. To prevent overfitting, an early stopping strategy is employed, with a patience value of 5 epochs which means training is halted if the validation loss does not improve for five consecutive epochs.

D. NEURAL NETWORK

The neural network for classification purposes consists of four dense layers with 512, 256, 128, and 64 units respectively, as shown in Part D of figure 2. A dropout layer is added before each of these dense layers and Relu is used as the activation function because it has been demonstrated empirically that it converges much more quickly and reliably than training a network with, for instance, the sigmoid activation function. The learned latent space from each person-specific CNN-VAE is stored and used as input to supervisely train this neural network where the target feature is the video ID, properly one-hot encoded. In detail, the same segmentation technique used with the CNN-VAE, including the same window length and shift, is used. Every latent space activated by each input tensor (32 × 32x5) of each trained person-specific CNN-VAE, for all the 40 videos for a participant were appended to a list. However, only the means of the n-dimensional latent space of each trained CNN-VAE model is considered as representative of the gaussian distribution and saved in this list. In line with the research hypothesis defined in section III, n is manipulated and the following dimensions are tested: 4, 8, 12, . . . , 92, 96, 100. The list was then shuffled and 70% of the data were used for training, 15% for validation, and 15% for testing. An early stopping strategy is also employed here to prevent overfitting of each model, and training is terminated after the validation accuracy does not improve for 5 consecutive epochs. These two architectures were implemented with the use of the Scikit-learn, Numpy, Pandas, Keras, and Tensorflow library packages.

E. MODELS EVALUATION

The evaluation of the learnt models, by training the autoencoder architecture (CNN-VAE) described in section III-C, is planned over two stages. The first stage includes the evaluation of the *reconstruction capacity* of the learned models against previously unseen testing data using the Structural Similarity Index (SSIM) and the Mean Squared Error (MSE) computed for the reconstructed topographic maps. In detail:

- SSIM - is a perceptual metric that measures how much image quality is lost as a result of processing, including data compression. It is an index of structural similarity (in the real range [0, 1] between two topographic maps (images). Values close to 1 indicate that the two topographic maps are very structurally similar, whereas values close to 0 indicate that the two images are exceptionally dissimilar and structurally different.
- MSE - it is defined as the mean (average) of the square of the difference between the actual and reconstructed values: the lower value indicates a better fit. In this case, the MSE involves the comparison, pixel by pixel, of the original and reconstructed topographic maps.

The second stage includes the evaluation of the *utility* of the learnt latent space of a trained CNN-VAE model for the classification of the different video categories, as part of the DEAP dataset, as described in section III-A. The impact of this latent space is evaluated by taking into account the accuracy, precision, recall, and F1-score of a neural network [39]:

- Accuracy - the formula for each target class is:

$$\text{ACCURACY} = \frac{TP_i + TN_i}{TP_i + FN_i + FP_i + TN_i} \tag{4}$$

  where TP, TN, FN, and FP stand for True Positive, True Negative, False Negative, and False Positive for i^{th} target class.
- Precision - it is determined as the sum of true positives across all classes divided by the sum of true positives and false positives across all classes.

$$\text{PRE} = \frac{\text{TP for all classes}}{\text{TP for all classes} + \text{FP for all classes}} \tag{5}$$

- Recall - it is determined as the sum of true positives across all classes divided by the sum of true positives and false negatives across all classes.

$$\text{REC} = \frac{\text{TP for all classes}}{\text{TP for all classes} + \text{FN for all classes}} \tag{6}$$

- F1-Score - it is the harmonic mean of two additional measures, namely precision and recall.

$$F_1 = 2 \cdot \frac{PRE \cdot REC}{PRE + REC} \tag{7}$$

IV. RESULTS & DISCUSSION

In this section, we discuss the results obtained by training the CNN-VAE models described in section III-C. In detail, findings are organised by the reconstruction capacity of these models, and the utility of their latent spaces to solve the video classification task.

A. RECONSTRUCTION CAPACITY OF CNN-VAE MODELS

Table 1 presents the SSIM and MSE scores of all the CNN-VAE models trained with different lengths of the time-slice and different latent space dimensions associated to only one participant. It is possible to note that the manipulation of these two parameters indeed has an effect on the performance of each CNN-VAE model. Here, for obvious space reasons, only the results associated to participant 1 are depicted. However, findings are consistent with these results across all the participants. Figure 5 reports the box-plots of the SSIM and MSE of all the CNN-VAE models for all

TABLE 1. An example of the SSIM and MSE scores of a one-person-specific convolutional variational autoencoder (CNN-VAE) of testing data for each time-slice length and latent space dimensions.

Latent space dimension	SSIM				MSE			
	Time-slice length							
	0.5sec	1sec	1.5sec	2sec	0.5sec	1sec	1.5sec	2
4	0.964	0.970	0.971	0.975	0.0003	0.0002	0.0002	0.000185
8	0.971	0.976	0.977	0.980	0.0002	0.0002	0.0002	0.000138
12	0.975	0.979	0.980	0.983	0.0001	0.0001	0.0001	0.000104
16	0.979	0.982	0.983	0.986	0.0001	0.0001	0.0001	8.65E-05
20	0.981	0.985	0.985	0.988	0.000109	9.21E-05	9.32E-05	7.27E-05
24	0.984	0.987	0.987	0.989	9.12E-05	7.76E-05	7.98E-05	6.20E-05
28	0.986	0.988	0.988	0.994	8.04E-05	6.86E-05	6.91E-05	3.63E-05
32	0.987	0.989	0.989	0.992	7.41E-05	6.49E-05	6.28E-05	4.74E-05
36	0.988	0.989	0.990	0.993	6.57E-05	6.10E-05	6.12E-05	4.15E-05
40	0.990	0.990	0.990	0.994	5.72E-05	5.57E-05	6.15E-05	3.70E-05
44	0.990	0.991	0.991	0.993	5.77E-05	5.43E-05	5.58E-05	3.86E-05
48	0.990	0.989	0.992	0.993	5.98E-05	6.16E-05	4.88E-05	3.89E-05
52	0.988	0.992	0.992	0.988	6.46E-05	4.65E-05	4.93E-05	7.08E-05
56	0.991	0.992	0.991	0.994	4.93E-05	4.79E-05	5.21E-05	3.37E-05
60	0.989	0.991	0.991	0.995	6.08E-05	5.00E-05	5.34E-05	3.05E-05
64	0.988	0.991	0.991	0.993	6.89E-05	4.98E-05	5.43E-05	4.09E-05
68	0.990	0.992	0.992	0.994	5.29E-05	4.75E-05	4.82E-05	3.22E-05
72	0.991	0.992	0.992	0.994	4.80E-05	4.85E-05	5.12E-05	3.25E-05
76	0.991	0.990	0.992	0.994	4.91E-05	5.70E-05	5.17E-05	3.19E-05
80	0.991	0.992	0.992	0.994	4.83E-05	4.87E-05	4.72E-05	3.37E-05
84	0.991	0.992	0.992	0.995	5.14E-05	4.31E-05	4.83E-05	3.02E-05
88	0.991	0.987	0.992	0.994	5.09E-05	7.32E-05	5.00E-05	3.24E-05
92	0.991	0.992	0.992	0.995	4.83E-05	4.84E-05	4.75E-05	2.92E-05
96	0.991	0.993	0.993	0.994	5.69E-05	4.14E-05	4.37E-05	3.40E-05
100	0.992	0.993	0.990	0.993	4.88E-05	4.20E-05	5.92E-05	4.19E-05

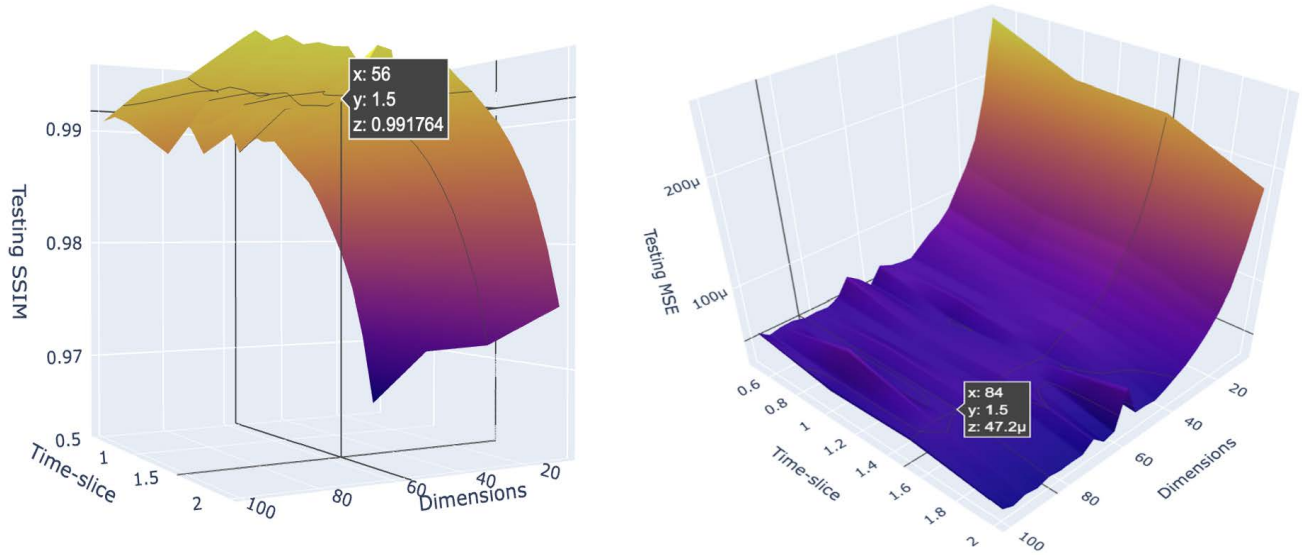

FIGURE 4. Graphical representation of the variation of the Structural Similarity Index (SSIM) and the Mean Squared Error (MSE) scores of one person-specific convolutional variational autoencoder (CNN-VAE) of the testing data as a consequence of the variation of the dimension of its latent space and the length of the EEG time-slice used for constructing the input EEG topographic head-maps.

participants, grouped by latent space dimension (x-axis) and time-slice length in seconds, for each of these dimensions.

Figure 4 depicts the variation of the Structural Similarity Index (SSIM) scores and the MSE scores to the manipulation

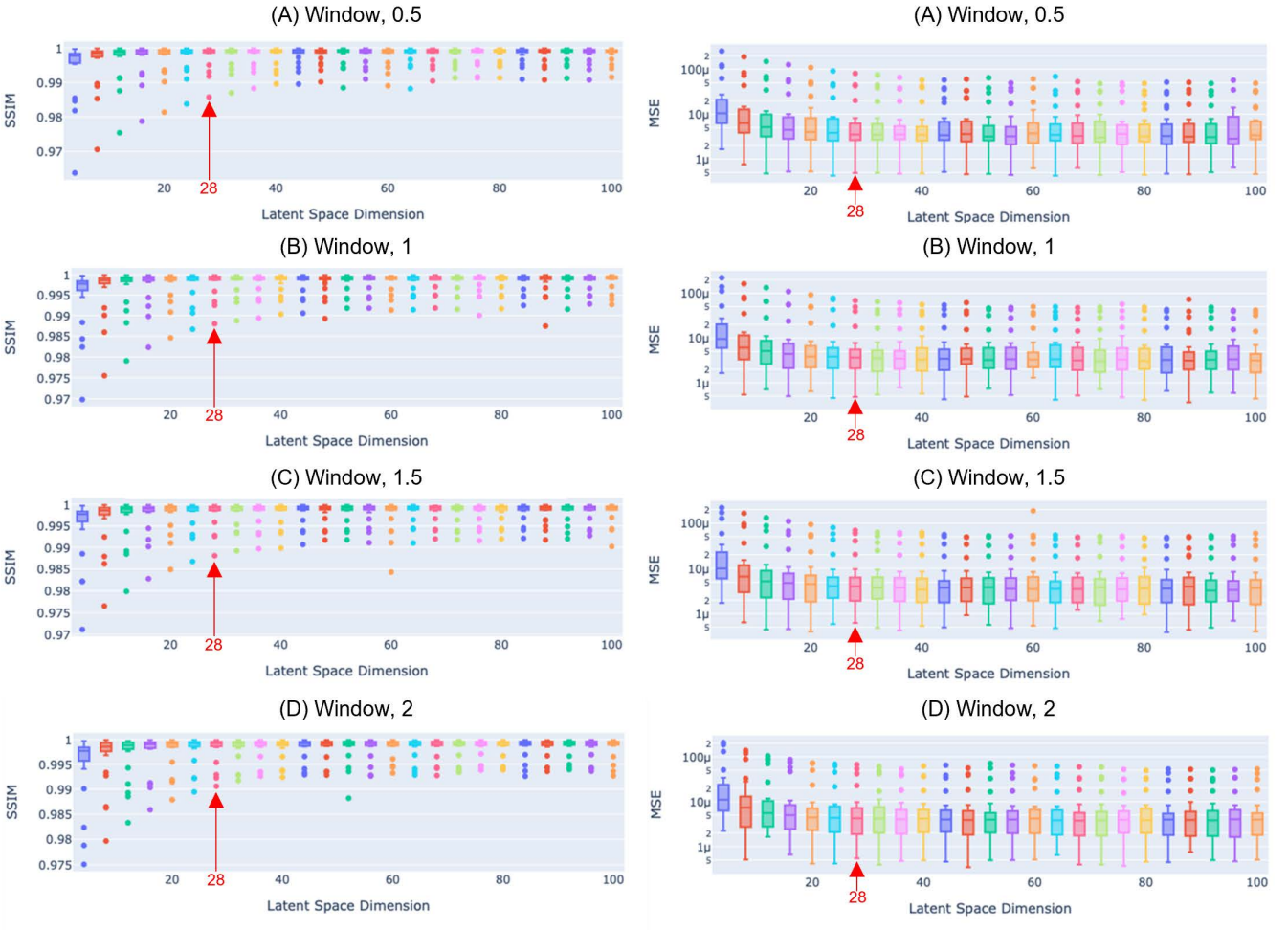

FIGURE 5. Box-plots of the SSIM and MSE scores of all the person-specific Convolutional Autoencoder (CNN-VAE) models, of the testing data, grouped by latent space dimension and time-slice length in seconds [A) 0.5 second B) 1 second C) 1.5 second D) 2 seconds].

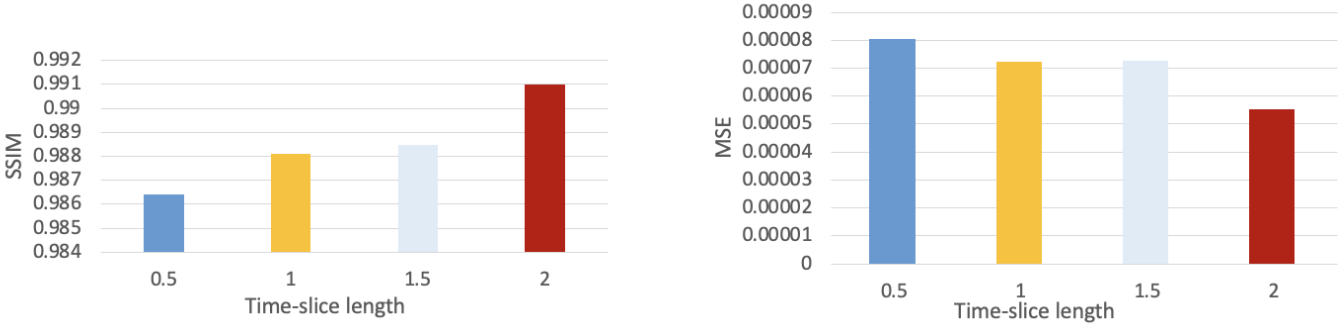

FIGURE 6. An example of the Average Structural Similarity Index (SSIM) and average Mean Squared Error (MSE) scores for one person-specific variational autoencoder of the testing topographic maps as a function of latent space dimension grouped by time-slice length.

of the length of the EEG time-slice, and the latent space dimension, for participant 1. Here, the CNN-VAE models trained with the EEG topographic head-maps generated with a higher time-slice length window outperformed the lower time-slice length in terms of the reconstruction capacity, as measured by the SSIM and MSE metrics. As the length of the window increases, the SSIM value approaches 1 and the MSE value approaches zero (figure 6). Subsequently, the average MSE and SSIM are plotted against each of the latent space dimensions shown in Figure 7. It can be seen that the latent space dimension has an effect on the performance of the variational autoencoders. As the latent space dimension

increases, SSIM approaches +1 and MSE approaches zero.

B. UTILITY OF THE LATENT SPACES FOR CLASSIFICATION

In this section, the utility of the latent space of each of the CNN-VAE models trained with EEG topographic head-maps for the classification of video categories, via the specific neural network (of section III-D) is presented. The performance of this network in the classification of video categories, given all four time-slice lengths (0.5, 1, 1.5, and 2 seconds) with different latent space dimensions (4 to 100 with a linear increment of 4) is tested by considering the accuracy, precision, recall, and F1-score metrics. Table 2 shows the accuracies of each trained neural network for the classification of video categories across all time-slices and latent space dimensions associated to only one participant. The result shows that accuracy with a 2 time-slice window length outperformed lower time-slice length windows. Similarly, increasing the latent space dimensions has an effect on the performance of the neural network shown in figure 9. Results are similar

TABLE 2. Neural network testing accuracy for classification of video categories with each time-slice length and all the latent space dimensions for a single participant.

Latent Dimension	Accuracy			
	Time-slice length			
	0.5sec	1sec	1.5sec	2sec
4	0.082	0.161	0.226	0.363
8	0.121	0.375	0.549	0.708
12	0.172	0.498	0.718	0.823
16	0.230	0.599	0.799	0.887
20	0.286	0.669	0.866	0.914
24	0.301	0.708	0.879	0.921
28	0.330	0.751	0.907	0.931
32	0.342	0.776	0.857	0.933
36	0.361	0.788	0.875	0.935
40	0.371	0.788	0.871	0.921
44	0.384	0.760	0.852	0.918
48	0.418	0.754	0.869	0.891
52	0.349	0.753	0.826	0.917
56	0.415	0.724	0.870	0.870
60	0.378	0.975	0.854	0.877
64	0.375	0.774	0.838	0.928
68	0.405	0.741	0.860	0.891
72	0.411	0.751	0.851	0.887
76	0.397	0.787	0.810	0.896
80	0.393	0.751	0.834	0.893
84	0.359	0.744	0.851	0.915
88	0.403	0.711	0.817	0.900
92	0.402	0.735	0.843	0.901
96	0.415	0.741	0.855	0.917
100	0.380	0.735	0.861	0.877

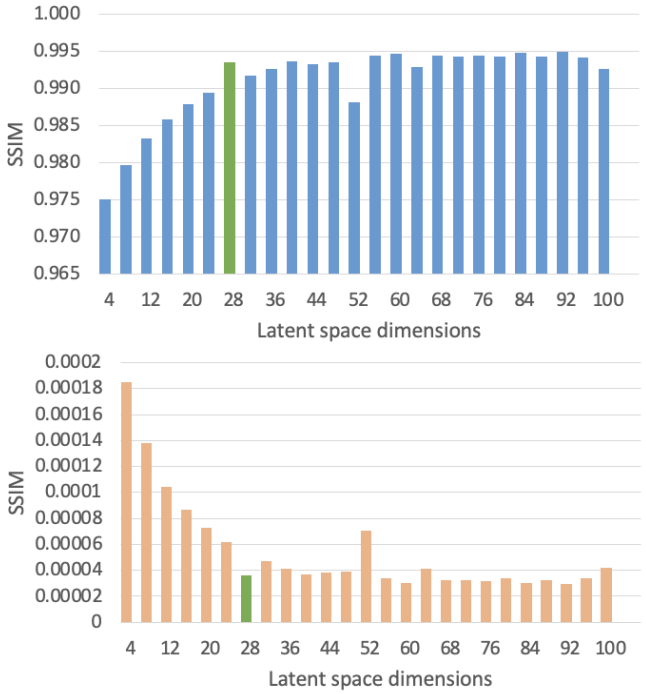

FIGURE 7. An example of the SSIM (top) and MSE (bottom) of testing data for each latent space dimension for a single participant.

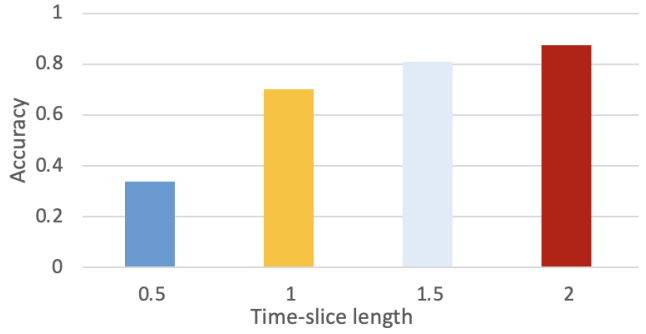

FIGURE 8. An example of average neural network testing accuracy for the classification of video labels as a function of each time-slice length for all latent space dimensions for a single participant.

and consistent across all the participants but are not reported for obvious space limits. It can also be seen in figure 11 reports the box-plots of the accuracy of all the neural network models for all participants, grouped by latent space dimension (x-axis) and time-slice length in seconds, for each of these dimensions.

Subsequently, the average accuracy are plotted against each time-slices length and each of the latent space dimensions as shown in figure 10 and figure 8. It can be clearly seen that a neural network trained with EEG topographic head-maps from higher time-slice lengths outperformed that trained with lower time-slice lengths. Similarly, the latent space dimension has an effect on the performance of the neural networks. These results apply to every model trained for each participant and are consistent.

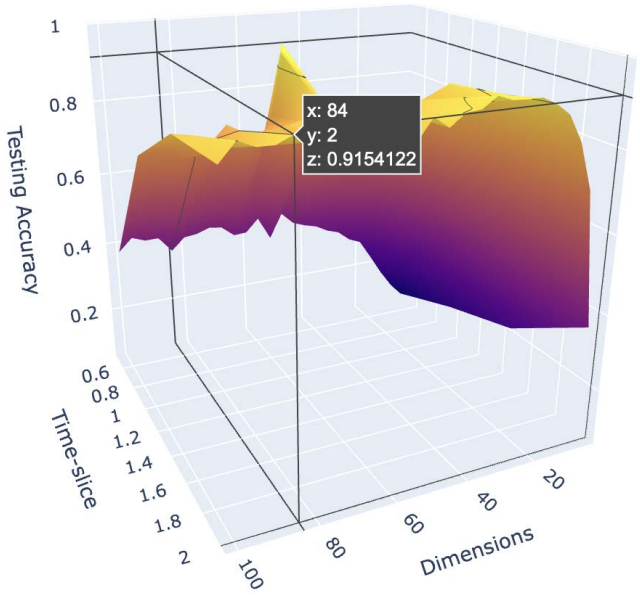

FIGURE 9. Graphical representation of the testing classification accuracy of video categories as a function of time-slice length and latent space dimension for a single participant.

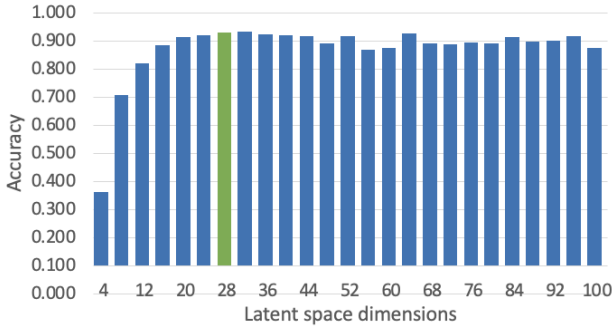

FIGURE 10. An example of neural network testing accuracy for the classification of video classes as a function of latent space dimension for a single participant.

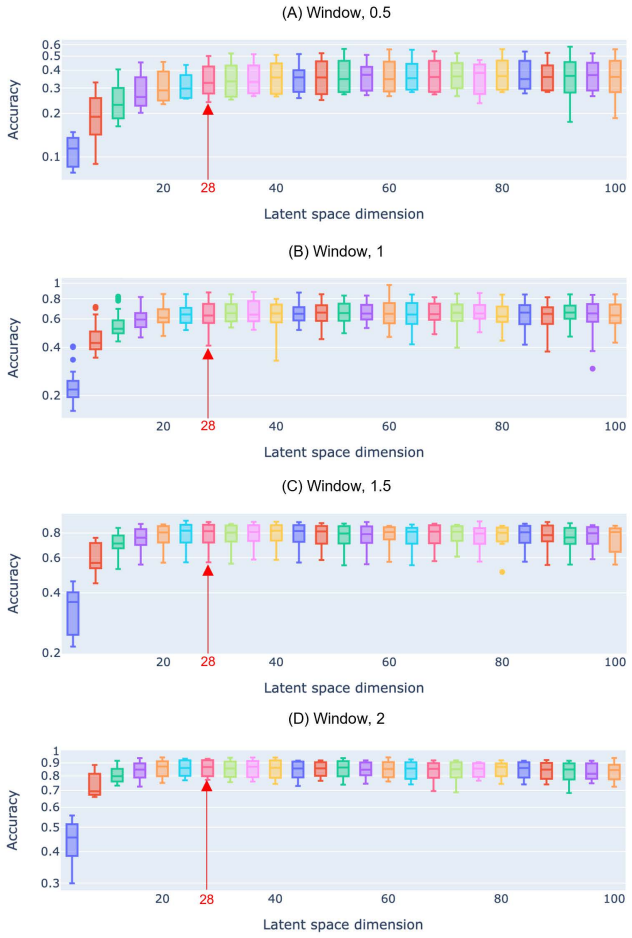

FIGURE 11. Box-plots of testing accuracy scores of the neural network models for all the participants grouped by latent space dimension and time-slice length in seconds [A) 0.5 second, B) 1 second, C) 1.5 second, D) 2 seconds].

The findings demonstrated that when a CNN-VAE is trained with topographic maps of shape (5, 32, 32) containing 5120 overall values, formed from 32 electrode values, for a 2 second window of cerebral activity, it is possible to reduce their dimensions by up to 99%, without losing salient information. In other words, from a tensor of 5120 values, each person-specific VAE could learn a latent space up to 28 means and 28 standard deviations without losing meaning, as assessed via a structural similarity index, and mean squared error between original and reconstructed tensors. These findings help test the initial hypothesis confirming that there exists a specific minimal size of the latent space of the designed VAE, that is effective both for maximal reconstruction capacity and for maximum accuracy when employed in a classification task.

Similar studies, although limited work with EEG signals and with 3-dimensional MR brains, have been conducted to determine the minimal latent space dimensions that maximise input reconstruction capacity and utility for classification tasks [10], [27]. The proposed pipeline for EEG data processing and high-level feature extraction via convolutional variational autoencoder, as designed in figure 2, has a number of advantages. First, transforming EEG signals from the time domain to the frequency domain can enable scholars to extract various types of brain waves (EEG bands) and thus relate them to various mental states. The variation of the mental state over time can serve as rich information about different aspects of human behavior that cannot be easily extracted from the time domain. Secondly, although this study has employed a dataset with 32 channels, the pipeline can be easily used with other amounts of electrodes. For example, [40], [41], and [42] have shown how topographic head maps of 32 × 32 pixels can be generated respectively with 19, 43, and 64 channels. Since also our pipeline generates topographic maps of 32 × 32 with 32 channels, then this means it can also generate the same size of maps from a larger number of electrodes. Thirdly, the autoencoder uses convolutional operations over input topographic maps, to learn salient

high-level features that are lower in dimension, and therefore these are more portable since they require a significantly less amount of digital memory to be stored. Additionally, this lower dimension contains the relevant and salient representations of EEG data that can be used for various purposes. These include the generation of synthetic EEG topographic head maps for data augmentation as well as employable in various classification tasks.

V. CONCLUSION

Researchers have designed and implemented different extensions of Variational Autoencoders with EEG signals for data augmentation, feature representation, and classification via latent space. However, research on the optimal dimension of the latent space of a VAE trained with EEG data is currently limited. The purpose of this study was to address this research challenge. An experiment has been conducted using an existing EEG dataset (DEAP) to establish the appropriate size of the latent space of person-specific VAEs that lead to the highest reconstruction capacity but also maximum utility in classification tasks. The dataset contains EEG data from 32 participants watching 60-second videos meant to trigger different human emotions, while 32 channels have been recorded using the 10-20 electrodes position standard. A sliding window technique has been used to isolate time windows from the EEG streams and create topology-preserving topographic head maps from them. In detail, a convolutional variational autoencoder architecture (CNN-VAE) has been designed and trained to learn high-level relevant representations from these head maps. Results show that manipulating both the size of these time windows and the latent space dimension, has an effect on the performance of resulting person-specific CNN-VAE models. The latent space generated from EEG head maps with a higher time slice length window outperformed the lower time slice in terms of the reconstruction capacity of the CNN-VAE. A second dense neural network has been devised to investigate the impact of such latent space on the classification of video categories for each participant. Similarly, as the latent space dimension increases, the learned latent space has an effect on the performance of this neural network. Future studies will include the interpretation of the latent space of the convolutional variational autoencoders, using principles from explainable artificial intelligence, in order to gauge further information and explain what has been learnt automatically [43]. For example, one way would be to visualise each layer of the decoder part of the variational autoencoder [44] to further understand which areas of the original topographic head maps have a greater impact on their reconstruction.

REFERENCES

[1] C. D. Binnie and P. F. Prior, "Electroencephalography," *J. Neurol. Neurosurg. Psychiatry*, vol. 57, no. 11, pp. 1308–1319, Nov. 1994.

[2] E. W. Anderson, G. A. Preston, and C. T. Silva, "Using Python for signal processing and visualization," *Comput. Sci. Eng.*, vol. 12, no. 4, pp. 90–95, Jul. 2010.

[3] Y. Sudaryat, J. Nurhadi, and R. Rahma, "Spectral topographic brain mapping in EEG recording for detecting reading attention in various science books," *J. Turkish Sci. Educ.*, vol. 16, no. 3, pp. 440–450, 2019.

[4] M. Taherisadr, M. Joneidi, and N. Rahnavard, "EEG signal dimensionality reduction and classification using tensor decomposition and deep convolutional neural networks," in *Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2019, pp. 1–6.

[5] A. M. Anwar and A. M. Eldeib, "EEG signal classification using convolutional neural networks on combined spatial and temporal dimensions for BCI systems," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 434–437.

[6] 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.

[7] J. Bornschein and Y. Bengio, "Reweighted wake-sleep," 2014, *arXiv:1406.2751*.

[8] S. M. Abdelfattah, G. M. Abdelrahman, and M. Wang, "Augmenting the size of EEG datasets using generative adversarial networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–6.

[9] J. F. Hwaidi and T. M. Chen, "A noise removal approach from EEG recordings based on variational autoencoders," in *Proc. 13th Int. Conf. Comput. Autom. Eng. (ICCAE)*, Mar. 2021, pp. 19–23.

[10] K. Li, J. Wang, S. Li, H. Yu, L. Zhu, J. Liu, and L. Wu, "Feature extraction and identification of Alzheimer's disease based on latent factor of multi-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1557–1567, 2021.

[11] A. Singh and T. Ogunfunmi, "An overview of variational autoencoders for source separation, finance, and bio-signal applications," *Entropy*, vol. 24, no. 1, p. 55, Dec. 2021.

[12] G. Krishna, C. Tran, M. Carnahan, and A. Tewfik, "Constrained variational autoencoder for improving EEG based speech recognition systems," 2020, *arXiv:2006.02902*.

[13] X. Li, Z. Zhao, D. Song, Y. Zhang, J. Pan, L. Wu, J. Huo, C. Niu, and D. Wang, "Latent factor decoding of multi-channel EEG for emotion recognition through autoencoder-like neural networks," *Frontiers Neurosci.*, vol. 14, p. 87, Mar. 2020.

[14] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1096–1103.

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

[16] V. D. Calhoun, T. Adali, G. D. Pearlson, and J. J. Pekar, "A method for making group inferences from functional mri data using independent component analysis," *Hum. Brain Mapping*, vol. 14, no. 3, pp. 140–151, 2001.

[17] M. Elsayed, K. S. Sim, and S. C. Tan, "Effective computational techniques for generating electroencephalogram data," in *Proc. Int. Conf. Adv. Sci. Eng. (ICOASE)*, Dec. 2020, pp. 1–5.

[18] Y. Luo, L.-Z. Zhu, Z.-Y. Wan, and B.-L. Lu, "Data augmentation for enhancing EEG-based emotion recognition with deep generative models," *J. Neural Eng.*, vol. 17, no. 5, Oct. 2020, Art. no. 056021.

[19] J. Yang, H. Yu, T. Shen, Y. Song, and Z. Chen, "4-class MI-EEG signal generation and recognition with CVAE-GAN," *Appl. Sci.*, vol. 11, no. 4, p. 1798, Feb. 2021.

[20] O. Ozdenizci and D. Erdogmus, "On the use of generative deep neural networks to synthesize artificial multichannel EEG signals," in *Proc. 10th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, May 2021, pp. 427–430.

[21] S. Liu, J. Liu, Q. Zhao, X. Cao, H. Li, D. Meng, H. Meng, and S. Liu, "Discovering influential factors in variational autoencoders," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107166.

[22] J. L. Hagad, T. Kimura, K.-I. Fukui, and M. Numao, "Learning subject-generalized topographical EEG embeddings using deep variational autoencoders and domain-adversarial regularization," *Sensors*, vol. 21, no. 5, p. 1792, Mar. 2021.

[23] O. Ozdenizci, Y. Wang, T. Koike-Akino, and D. Erdogmus, "Transfer learning in brain-computer interfaces with adversarial variational autoencoders," in *Proc. 9th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Mar. 2019, pp. 207–210.

[24] J. F. Hwaidi and T. M. Chen, "A novel KOSFS feature selection algorithm for EEG signals," in *Proc. IEEE EUROCON 19th Int. Conf. Smart Technol.*, Jul. 2021, pp. 265–268.

[25] L. Bi, J. Zhang, and J. Lian, "EEG-based adaptive driver-vehicle interface using variational autoencoder and PI-TSVM," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 2025–2033, Oct. 2019.

[26] Z. Chen, N. Ono, M. Altaf-Ul-Amin, S. Kanaya, and M. Huang, "IVAE: An improved deep learning structure for EEG signal characterization and reconstruction," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 1909–1913.
[27] C. Vogelsanger and C. Federau, "Latent space analysis of VAE and intro-VAE applied to 3-dimensional MR brain volumes of multiple sclerosis, leukoencephalopathy, and healthy patients," 2021, *arXiv:2101.06772*.
[28] A. Vereshchaka, F. Yang, A. Suresh, I. L. Olokodana, and W. Dong, "Predicting cognitive control in older adults using deep learning and EEG data," in *Proc. Int. Conf. Social Comput., Behav.-Cultural Modeling Predict. Behav. Represent. Modeling Simulation (SBP-BRiMS)*, Washington, DC, USA, Oct. 2020, pp. 19–22.
[29] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, and M. Shah, "Brain2Image: Converting brain signals into images," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1809–1817.
[30] D. Ayata, Y. Yaslan, and M. Kamasak, "Multi channel brain EEG signals based emotional arousal classification with unsupervised feature learning using autoencoders," in *Proc. 25th Signal Process. Commun. Appl. Conf. (SIU)*, May 2017, pp. 1–4.
[31] A. M. Abdelhameed and M. Bayoumi, "An efficient deep learning system for epileptic seizure prediction," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2021, pp. 1–5.
[32] A. M. Abdelhameed and M. Bayoumi, "Semi-supervised EEG signals classification system for epileptic seizure detection," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1922–1926, Dec. 2019.
[33] H. T. D. Tran, "Eeg-based person authentication using variational universal background model," M.S. thesis, Univ. Canberra, Canberra, ACT, Australia, 2019.
[34] X. Li, Z. Zhao, D. Song, Y. Zhang, C. Niu, J. Zhang, J. Huo, and J. Li, "Variational autoencoder based latent factor decoding of multichannel EEG for emotion recognition," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 684–687.
[35] S. Koelstra, "Deap: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2011.
[36] M. Abo-Zahhad, S. M. Ahmed, and S. N. Abbas, "A new EEG acquisition protocol for biometric identification using eye blinking signals," *Int. J. Intell. Syst. Appl.*, vol. 7, no. 6, p. 48, 2015.
[37] A. Kristiadi, "Variational autoencoder: Intuition and implementation," Agustinus Kristiadi's Blog, Univ. Tubingen, Germany, 2020. Accessed: Dec. 10, 2016. [Online]. Available: https://agustinus.kristia.de/techblog/2016/12/10/variational-autoencoder/
[38] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, 2014, pp. 1–10.
[39] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *Int. J. Data Mining Knowl. Manag. Process*, vol. 5, no. 2, pp. 1–11, Mar. 2015.
[40] N. Wagh, J. Wei, S. Rawal, B. Berry, L. Barnard, B. Brinkmann, G. Worrell, D. Jones, and Y. Varatharajah, "Domain-guided self-supervision of EEG data improves downstream classification performance and generalizability," in *Proc. Mach. Learn. Health*, 2021, pp. 130–142.
[41] S. Zhang, X. Mao, L. Sun, and Y. Yang, "EEG data augmentation for personal identification using SF-GAN," in *Proc. 3rd Int. Conf. Comput. Vis., Image Deep Learn. Int. Conf. Comput. Eng. Appl. (CVIDL ICCEA)*, May 2022, pp. 1–6.
[42] S. Rayatdoost, D. Rudrauf, and M. Soleymani, "Expression-guided EEG representation learning for emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3222–3226.
[43] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Inf. Fusion*, vol. 76, pp. 89–106, Dec. 2021.
[44] G. Vilone and L. Longo, "Classification of explainable artificial intelligence methods through their output formats," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 3, pp. 615–661, Aug. 2021.

TAUFIQUE AHMED received the B.E. and M.Tech. degrees in computer science engineering from Visvesvaraya Technological University, Belagavi, India, in 2009 and 2013, respectively. He is currently pursuing the Ph.D. degree with the Artificial Intelligence and Cognitive Load Research Laboratory, School of Computer Science, Technological University Dublin, Ireland. He was a Lecturer at the Department of Computer Science and Engineering, Anjuman Institute of Technology and Management, Bhatkal, India, from 2013 to 2021. His main research interests include cognitive load modeling, machine learning, deep learning, and natural language processing.

LUCA LONGO received the bachelor's and master's degrees in computer science, statistics, and health informatics and the doctoral degree in artificial intelligence from Trinity College Dublin. He is a Leader with the Artificial Intelligence and Cognitive Load Research Laboratories, with his team of doctoral and postdoctoral students, he conducts fundamental research in explainable artificial intelligence, defeasible reasoning, and non-monotonic argumentation. He also performs applied research in machine learning and predictive data analytics, mainly applied to the problem of mental workload modeling. He is actively engaged in the dissemination of scientific material to the public contributing to the non-profit TED Organization.