2022

On the Dimensionality and Utility of Convolutional Autoencoder's Latent Space Trained with Topology-Preserving Spectral EEG Head-Maps

Arjun Vinayak Chikkankod
Technological University Dublin, arjunvinayak.chikkankod@tudublin.ie

Luca Longo
Technological University Dublin, luca.longo@tudublin.ie

Recommended Citation

Chikkankod, A.V., & Longo, L. (2022). On the Dimensionality and Utility of Convolutional Autoencoder's Latent Space Trained with Topology-Preserving Spectral EEG Head-Maps. *Machine Learning Knowledge Extraction, vol. 4, no. 4, pg. 1042-1064. https://doi.org/10.3390/make4040053*

machine learning & knowledge extraction

MDPI

Article

On the Dimensionality and Utility of Convolutional Autoencoder's Latent Space Trained with Topology-Preserving Spectral EEG Head-Maps

Arjun Vinayak Chikkankod * and Luca Longo

Artificial Intelligence and Cognitive Load Lab, The Applied Intelligence Research Centre, School of Computer Science, Technological University Dublin (TU Dublin), D07 EWV4 Dublin, Ireland

* Correspondence: arjunvinayak.chikkankod@tudublin.ie

Abstract: Electroencephalography (EEG) signals can be analyzed in the temporal, spatial, or frequency domains. Noise and artifacts during the data acquisition phase contaminate these signals adding difficulties in their analysis. Techniques such as Independent Component Analysis (ICA) require human intervention to remove noise and artifacts. Autoencoders have automatized artifact detection and removal by representing inputs in a lower dimensional latent space. However, little research is devoted to understanding the minimum dimension of such latent space that allows meaningful input reconstruction. Person-specific convolutional autoencoders are designed by manipulating the size of their latent space. A sliding window technique with overlapping is employed to segment varied-sized windows. Five topographic head-maps are formed in the frequency domain for each window. The latent space of autoencoders is assessed using the input reconstruction capacity and classification utility. Findings indicate that the minimal latent space dimension is 25% of the size of the topographic maps for achieving maximum reconstruction capacity and maximizing classification accuracy, which is achieved with a window length of at least 1 s and a shift of 125 ms, using the 128 Hz sampling rate. This research contributes to the body of knowledge with an architectural pipeline for eliminating redundant EEG data while preserving relevant features with deep autoencoders.

Keywords: electroencephalography; latent space analysis; sliding windowing; convolutional autoencoders; automatic feature extraction; dense neural network

Citation: Chikkankod, A.V.; Longo, L. On the Dimensionality and Utility of Convolutional Autoencoder's Latent Space Trained with Topology-Preserving Spectral EEG Head-Maps. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 1042–1064. https://doi.org/10.3390/make4040053

Academic Editor: Javier Del Ser Lorente

Received: 24 September 2022
Accepted: 14 November 2022
Published: 18 November 2022

1. Introduction

In the brain, multiple sources operate synchronously to carry out a specific mental task. For instance, the brain's occipital lobe perceives visual data, the temporal lobe sense auditory signals, the parietal lobe gathers sensory data, the limbic lobe generates emotion, the insula processes pain, and the frontal lobe makes decisions [1]. The brain executes lobular activities in the span of milliseconds. Electroencephalography (EEG) is the most frequently used physiological indicator to understand brain functioning and behavior. EEG measures the voltage potentials of the neuronal excitations via electrodes mounted on the head with a greater temporal resolution and precision, suggested to match the speed of cognition [2–4]. EEG data can be analyzed in the temporal, spatial, or frequency domains. However, noise and artifacts from eye and muscle movements during the data acquisition phase contaminate EEG signals adding difficulties in their analysis.

Optimal EEG features are devoid of the noise and artifacts that capture the variability of entire input EEG signals to the maximum possible extent by using smaller dimensions. Hence there is a strong need to automate the extraction of optimal EEG features from multidimensional EEG signals.

Several studies have used hand-crafted features and traditional machine learning algorithms to discover salient features from EEG data. Hand-engineered feature selection can give an optimal solution. However, they are highly application-specific, for

example, removal of outlying channels and EEG segments corrupted with noise and artifacts by visual inspection [5]. Moreover, hand-crafted features learn from trial and error, which consumes time and human effort [6]. Principal Component Analysis (PCA) is a dimensionality-reduction algorithm that projects higher dimensional feature vectors into lower-dimensional orthogonal features [7,8]. Orthogonal features are independent variables that share no correlation, thus contributing to decision-making. Orthogonality in the feature space produces the optimal features that explain the variability of the entire input in the best possible way. However, the projected lower-dimensional feature space is linear, making the principal components less interpretable [9,10]. Independent Component Analysis (ICA) isolates noise and artifacts from the input source by decomposing them into independent components from different sources [9,11]. Noise and artifacts appear as high-frequency spikes, low-frequency drifts, or periodic fluctuations and can thus be identified and removed. ICA algorithms such as InfomaxICA and FastICA work better with stationary signals [12,13]. Adaptive Mixture ICA is suitable for non-stationary data but is computationally expensive. Moreover, its effectiveness depends on the EEG Signal's Window Length (WL) and the number of channels [14,15]. Traditional methods such as PCA and ICA do not improve with large EEG datasets [16]. Thus recent studies use Convolutional Neural network (CNN)-based methods and Autoencoders to extract useful features. Features selected from CNN-based methods improve the model's accuracy compared to traditional algorithms on large datasets. However, the drawback is that the quality of the automatically learned features by CNN layers is not examined [17,18]. Autoencoder examines the quality of the extracted features by reconstructing input from the latent space and validating with the true inputs [19]. However, the optimal dimension of the Autoencoder's latent space is unknown in the existing literature. Exploratory analysis of the latent space dimension's power must be carried out to establish Autoencoder's best latent space dimension.

Our study aims to identify the optimal window length (WL) and window shift (WS) of the sliding windowing from EEG signals that leads to an optimal latent space (LS) learned by training person-specific Autoencoders. The fixed-size sliding windowing facilitates the generation of multiple input instances of specified window length by linearly traversing the EEG data with a specified window shift [20]. Person-specific Autoencoder captures the features from an individual subject's EEG data as the brain wave response of every individual are unique and carry higher inter-subject variability for rest, task, and activity evoked brain states [21]. Furthermore, the significance of the learned latent space is examined by using reconstruction measures, including the Structural Similarity Index Measure (SSIM), the Mean Squared Error (MSE), the Normalised Root Mean Squared Error (NRMSE), and the Peak Signal-to-Noise Ratio (PSNR). Specifically, a classification task's utility is investigated to validate the learned latent space, including the accuracy and F1-score measures. The above research objectives are aimed at answering the following research question (RQ):

> RQ: *What are the optimal window length and window shift to segment continuous EEG signals that leads to the formation of a latent space in person-specific Convolutional Autoencoder that leads to maximum reconstruction capacity and maximum utility in classification tasks?*

The remainder of the article is organized as follows. A background on EEG feature space with traditional pre-processing techniques is described in Section 2. Section 3 presents the design of our experiment to construct person-specific convolutional autoencoders that solve the research question along with the evaluation metrics. Section 4 presents the findings, followed by a discussion in Section 5. Eventually, Section 6 summarizes this study, highlighting its contribution to the body of knowledge and delineating future areas of work.

2. Related Work

Electroencephalography (EEG) is a widely used physiological technique that is easy to operate and economical, which aims to measure the electrical activity of the neurons in the brain propagated at the scalp level [9]. Thus EEG analysis has been extensively used in diverse application areas such as epilepsy seizure onset prediction [22], brain-computer interface (BCI) [23–25], emotion recognition [26,27], driver distraction [28], mental workload measurement [29], and many other neurological disorder diagnoses [30,31]. However, EEG is a multidimensional, non-stationary signal with poor signal-to-noise ratio characteristics [32]. EEG signals must be pre-processed to gain better insight into the data and perform better analysis.

EEG signals can be analyzed in temporal, frequency, or spatial domains. The temporal domain gives rhythmic voltage fluctuations at every time point for all the recorded channels [2,32]. Temporal analysis methods such as the Time Domain Parameters (TDP) display nominal computational complexity while maintaining acceptable outcomes for classification tasks [33]. However, it is challenging to identify noise and artifacts in the time domain as they exhibit randomness over frequencies. In contrast, noise identification is easier when analyzing EEG data in the frequency domain. Noise is often spread across entire frequency bands, and the noise amplitude is negligible compared to the amplitude of the actual signal itself, thus more identifiable [2,32]. Additionally, in the frequency domain, EEG data can be analyzed using an amplitude spectrum that separates signals into frequency bands such as delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz) and gamma (30–45 Hz) bands, usually employing Fourier transformation [34]. Eventually, in the spatial domain, the locations of the electrodes can also give important information, thus enriching the EEG data analysis [34,35]. For example, the EEG signal recorded at a specific location could be de-noised considering the signals coming from neighboring electrodes. Unfortunately, very few studies have focused on analyzing EEG data in all three domains.

Several approaches exist to reduce noise and redundant features in EEG signals. PCA tries to identify the subspace to represent the EEG data by eliminating linearly dependent features [9,36,37]. Linear Discriminant Analysis (LDA) is a dimensionality reduction method that maximizes inter-cluster spread and minimizes intra-cluster spread among data points [9,36]. ICA is a linear decomposition technique that finds a new basis to represent the data, solving the source separation problem [38]. A newfound base separates independent components from the mixed signals and noise. Artifacts in the independent components are removed by back-projection [9,36]. Common Spatial Pattern (CSP) algorithm and their extensions [34,35,39] alleviate the adverse effect of noise, signal artifacts, and non-stationary signal behavior by splitting EEG signals into additive segments that have the most dissimilarities in the variance between two adjacent windows [40]. All these techniques decompose EEG signals into different representations, sometimes of lower dimensions. These representations are either automatically or manually inspected for noise's presence, and those containing noise are removed, retaining only the relevant ones that are used to reconstruct EEG signals. Data from sliding windowing further increases the efficiency of the previously listed techniques in extracting relevant features from multidimensional EEG data, thereby becoming an effective tool for neural signal analysis [20,41]. A fixed-length window will move along the time dimension of EEG signals with a predetermined shift in the sliding window technique.

Very often, the relevant features extracted using the previous techniques are fed to Machine Learning (ML) algorithms such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forests (RF), or Gaussian Mixture Models (GMM) for prediction or classification tasks [9,42–46]. These classifiers have proven to be fast and highly accurate for small and medium-size datasets [42]. However, their performance does not improve significantly for large datasets. On the contrary, the performance of Deep Learning (DL) algorithms improves with a large dataset [47,48]. DL algorithms have also been employed in the EEG pre-processing phase to automatically denoise EEG signals. These algorithms

aim to discover vital features from EEG signals, thus automatizing the manual analysis, often performed by humans, in the temporal, frequency, or spatial domains [18,26,30]. For example, Convolutional Neural Networks (CNN) have been used to learn the essential EEG features, improving models' performance compared to traditional ML methods [13,17]. However, the quality of the selected features is often neither explained nor validated using evaluation metrics, but 'trusted' by researchers and engineers. Fortunately, within the broader field of Deep Learning, an unsupervised technique can validate extracted features using reconstruction metrics. Autoencoders are unsupervised neural networks that encode the input into a lower dimensional latent space (LS), which is supposed to retain the essential features of the input and discard the irrelevant ones, such as redundant features and noise [49]. The latent space is then decoded to reconstruct the input. This step is crucial as it represents a straightforward way of validating the quality of latent space by evaluating its capacity to reconstruct the input with which the model was trained successfully. Autoencoders are widely used in domains such as image denoising, compression, reconstruction, dimensionality reduction, and visualization [4,13,50]. Recently, Autoencoders have been used for neural signal analyses [17,19].

However, an investigation and exploratory analysis of the optimal dimension of the Autoencoder's latent space is still an open question. In particular to EEG signals, it is essential to understand: (a) what is the ideal size of EEG windows that can be used to construct single input instances for Autoencoders? (b) what is the effect of different window time shifts? (c) what is the smallest latent space that can be constructed from the Autoencoders without losing vital information?

3. Materials and Methods

An empirical study with secondary data was designed and implemented to answer the research question. The research hypothesis was defined as follows:

> *IF a sliding window technique is used to segment multichannel EEG signals into windows, AND topographic head-maps are formed from each window, which is used to train a Convolutional Autoencoder (ConvAE) for reducing their dimensionality*
> *THEN there exists an optimal window length (WL) AND window shift (WS) combination that leads to the formation of a minimal latent space (LS) for ConvAE that maximizes the mean reconstruction capacity of the input topographic maps, AND that has maximal mean utility in a classification task.*

While on the one hand, the mean reconstruction capacity is evaluated by employing the Structural Similarity Index Measure (SSIM), the Mean Squared Error (MSE) and its normalized version (NRMSE), as well as the Peak Signal-to-Noise Ratio (PSNR) of the input topographic maps against the reconstructed ones, on the other hand, the mean utility of the latent space is measured by accuracy and F1-score for a chosen classification task. In order to test the research hypothesis, a set of phases has been designed, as illustrated in Figure 1, and explained in the following subsections.

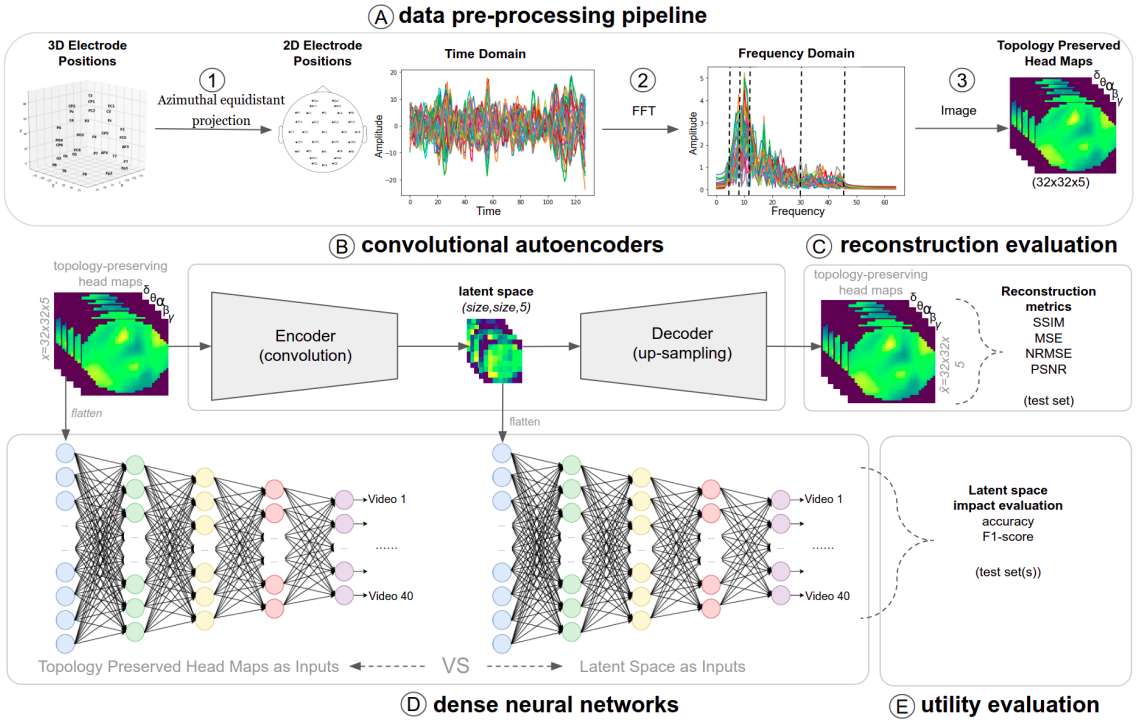

Figure 1. (**A**) Data pre-processing pipeline gives topographic head-maps (TPHM) as output from 32 channel EEG signals (**B**) ConvAE for obtaining the latent space (**C**) Reconstruction metrics to evaluate the latent space (LS) quality (**D**) DNN with LS and TPHM as inputs (**E**) Utility (U) to measure the model's performance.

3.1. DEAP Dataset

DEAP is a popular dataset of multichannel EEG data and peripheral physiological data from 32 subjects [51]. Signals were recorded using the Biosemi ActiveTwo system while each subject watched 40 one-minute-long music videos. Every video had the artist's name, title, and URL comprising various emotion tags such as happy, fun, sad, exciting, pleasure, joy, and much more. DEAP is a benchmark dataset for video-induced emotion research. Three-second baseline signals were added at the beginning of every one-minute EEG recording. DEAP has an overall 40 channels, out of which 32 were EEG signals, and the remaining 8 channels were peripheral signals. The 32 AgCl electrodes were mounted on the scalp as per the international 10-20 standard and using the following channels: *Fp1*, *AF3*, *F3*, *F7*, *FC5*, *FC1*, *C3*, *T7*, *CP5*, *CP1*, *P3*, *P7*, *PO3*, *O1*, *Oz*, *Pz*, *Fp2*, *AF4*, *Fz*, *F4*, *F8*, *FC6*, *FC2*, *Cz*, *C4*, *T8*, *CP6*, *CP2*, *P4*, *P8*, *PO4*, *O2*.

3.2. Data Pre-Processing

The experiment's first phase includes signal pre-processing (phase A of Figure 1). EEG signals were captured with a sampling rate of 512 Hz, which was down-sampled to 128 Hz [51]. Electro-oculogram (EOG) artifacts resulting from eye blinking were removed as part of the data-cleaning process. A band-pass frequency filter from 0.5–45.0 Hz was applied to the recorded EEG signals. Three-second baseline data were also removed for every EEG signal associated with each video. The electrodes positioned on the human scalp in the 3D space were mapped onto a 2D Cartesian plane using Azimuthal equidistant projection (polar projection) (1 of Figure 1). The EEG signals gathered from each participant and each video was sliced into windows with size 0.5 s, 1 s, 1.5 s, and 2 s and with window shift of 125 ms, 250 ms, and 500 ms respectively. As anticipated in the research hypothesis, the possible window length and shift configurations will be empirically evaluated. Noticeably, these windows overlap significantly, and the rationale was to have a considerable amount of windows that could be used as input to the subsequent deep learning phase. The next

step in the pre-processing pipeline was to transform the EEG signals into the frequency domain by employing the Fast Fourier Transform (FFT) (2 of Figure 1). The transformed data in the frequency domain were separated into the five EEG bands: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–45 Hz) [34].

The final step was to generate topographic head-maps (TPHM) using the information from the five EEG bands. A mesh grid of size 32×32 was created for each band by using the 2D electrode positions on the grid. The piecewise cubic interpolation method was used to fill values along the 2D grid surface with 32 electrode positions (3 of Figure 1). The five head-maps (for delta, theta, alpha, beta, and gamma) were joined into a tensor of shape $32 \times 32 \times 5$ which represented the input to a Convolutional Autoencoder (ConvAE) (phase B of Figure 1) and a Dense Neural Network (DNN) (phase D of Figure 1), both described in the following sections. In summary, the shape of the data associated with each participant was of the order $40 \times 32 \times 128 \times 60$ that corresponded to the number of videos, the number of channels, sampling frequency, and video duration. These data, as mentioned previously, were further segmented with an overlapping sliding window technique. The window shift (WS) determined the extent of the overlapping among consecutive frames. 0.5 s window length achieved an overlap of 75%, 50%, and 0% (no overlap) from shifts of 125 ms, 250 ms, and 500 ms. With 1 s window length, an overlap of 87.5%, 75%, and 50% was achieved. Using 1.5 s window length, an overlap of 91.7%, 83.3%, and 66.7% was noted. Finally, with 2 s window length, an overlap of 93.7%, 87.5%, and 75% was achieved with the same shifts of 125 ms, 250 ms, and 500 ms respectively. Topographic maps were created for each window. Thus the size of the resulting datasets changed based on the chosen window length and window shift configurations. In detail, Table 1 shows the number of windows, each useful for generating a $32 \times 32 \times 5$ input tensor (5 topographic head-maps, one for each EEG band) generated for each of the above configurations.

Table 1. Amount of EEG windows extracted from EEG data per video (60 s) and in total (for 40 videos) as a function of window length (in s) and window shift (in ms) for each participant.

Window Length (s)	Window Shift (ms)	Amount (in 1 Video)	Amount (Total)
0.5	125	477	19,080
0.5	250	239	9560
0.5	500	120	4800
1.0	125	473	18,920
1.0	250	237	9480
1.0	500	119	4760
1.5	125	469	18,760
1.5	250	235	9400
1.5	500	118	4720
2.0	125	465	18,600
2.0	250	233	9320
2.0	500	117	4680

3.3. Convolutional Autoencoders (ConvAE)

Autoencoders learn optimal features from unlabelled input data without supervision. The learning algorithm performs backpropagation by assigning input data as target values, which means $y^i = x^i$ where x and y represent input and output, respectively, for an ith training example [52]. The model h (1) learns an approximation to the identity function with weights W and bias b.

$$h_{W,b}(x) \approx x \tag{1}$$

Thus the generated output is similar to the input. An autoencoder consists of an encoder–decoder block. The output of the encoder is the latent space (LS) which preserves optimal features usually in a lower dimensional space, thus capturing the essence of input. An autoencoder reconstructs the input by retaining essential features in the latent space,

thereby eliminating redundant ones. The reconstructed output has the same shape as the input. The holdout validation approach was used to partition the generated topographic maps from each of the datasets (Table 1) into train, validation, and test sets, respectively containing 75%, 12.5%, and 12.5% of the available data.

The ConvAE (phase B of Figure 1) consisted of multiple convolutional and max-pooling layers. The initial convolutional layer had 64 kernel units. The kernels were doubled for every subsequent convolutional layer to compensate for the reduced input size at every passing max-pooling layer (Figure 2). The decoder was symmetrical to the encoder with an equal number of layers. However, the kernels were halved at the subsequent convolutional layer. Each convolutional layer used 3×3 kernels to convolve on each level (5 EEG bands) of the input topographic head-map tensor ($32 \times 32 \times 5$). The rationale behind the kernel size is that small-sized kernels are cost-effective, and odd-sized kernels possess symmetric properties that make convolution without distortion [53]. Every convolutional step used zero padding to prevent shrinkage of the image dimensions. Padding ensured the input size remained the same at every convolutional layer. ReLU, a non-linear activation function, was used in every layer:

$$f(z) = \begin{cases} z, & \text{if } z > 0. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where $z = Wx + b$.

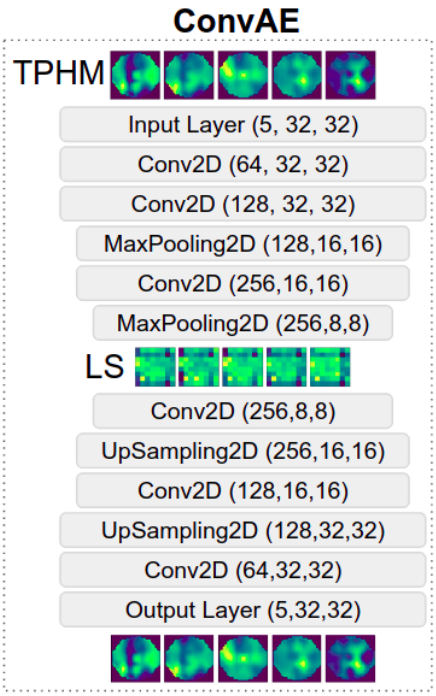

Figure 2. Convolutional Autoencoder's (ConvAE) architecture for learning optimal latent space (LS) from topology-preserving head-maps (TPHM) for delta, theta, alpha, beta, and gamma EEG bands.

The Autoencoder used Adam as an optimization algorithm where the optimal learning rate was determined by performing hyperparameter tuning. The Autoencoder's loss was measured using the Mean Squared Error (MSE), which is given by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(X^i - \hat{X}^i)^2 \tag{3}$$

where n represents the number of predicted data points. The latent space (LS) was set to be $(2, 2, 5)$, $(4, 4, 5)$, $(8, 8, 5)$, and $(16, 16, 5)$ respectively. Convolutional autoencoders

were person-specific, meaning that only the data from a single participant were considered for training a model. Thus 32 models, one for each participant and each configuration of Table 1 were generated. The Autoencoder was trained for 100 epochs by incorporating an early stopping mechanism.

3.4. Reconstruction Evaluation Metrics

Structural Similarity Index (SSIM), Mean Squared Error (MSE), Normalised Root Mean Squared Error (NRMSE), and Peak Signal-to-noise Ratio (PSNR) were used to validate the effectiveness of the person-specific autoencoder models and their reconstruction capacity (phase C of Figure 1). SSIM measures the similarity between two topographic head-maps x and y:

$$\text{SSIM} = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{4}$$

where μ_x, μ_y, σ_x, and σ_y are mean and variance of x and y respectively, and c_1 and c_2 are the stabilization constants of division. SSIM ranges between 0 and 1, where 1 means the perfect match between the original and reconstructed topographic head-maps. SSIM measures degradation in the reconstructed input resulting from data compression, which is based on perception and saliency error properties [54].

MSE gives the cumulative squared difference between the original and reconstructed topographic head-maps (Equation (3)). MSE gives absolute error [54] and is strictly positive: the lower the MSE, minimal is the reconstruction error. RMSE measures the dissimilarity between reconstructed and true pixel values from reconstructed and original topographic head maps. RMSE is normalized, which facilitates the comparison of two topographic maps with different scales. NRMSE with mean $\bar{y}$ for the measured data is defined by

$$\text{NRMSE} = \frac{\text{RMSE}}{\bar{y}} \tag{5}$$

$$\text{RMSE} = \sqrt{\text{MSE}(\hat{X})} \tag{6}$$

Similar to MSE, a lower NRMSE indicates more similarity between the two topographic head-maps.

PSNR gives a peak signal-to-noise ratio for two different topographic head-maps (Equation (7)) [54]. PSNR is measured in decibels where higher values indicate better reconstruction quality.

$$\text{PSNR} = 10\log_{10}\frac{R^2}{\text{MSE}} \tag{7}$$

R corresponds to the maximum amplitude variation in the input. The reconstruction quality is measured based on perception, saliency, and absolute error properties.

3.5. Classification

A neural network was designed for a classification task to demonstrate the utility of the latent space, learned by training the ConvAE (phase D of Figure 1). The goal was to train a new predictive model using latent space as input and a video category as the target feature (DEAP dataset, Section 3.1). The model was trained using a fully connected Dense Neural Network (DNN) with input, hidden, and output layers (Figure 3) [55]. The input was flattened and passed to the first dense layer with 512 units and the ReLU activation function. The number of units in the subsequent hidden layers was decreased by a power of 2. Each hidden layer obtained its input from the previous layer, computed weight metrics and bias typically using ReLU, a non-linear activation function, and transferred its output to the next layer. Every hidden layer was interleaved with a Dropout layer with a dropout

rate of 0.1. The output layer had 40 units with a softmax activation function. The Softmax function was used at the output layer to work with a multiclass classifier.

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{8}$$

σ is the softmax function, z is the input vector, e^{z_i} and e^{z_j} are the standard exponentiation functions, and k represents the number of classes in the multiclass classification.

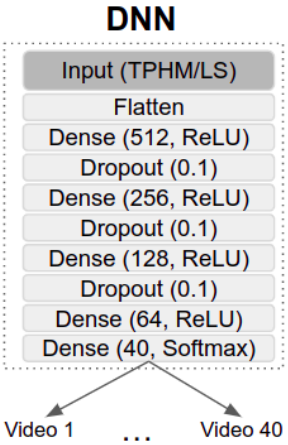

Figure 3. A fully connected Dense Neural Network (DNN) architecture for predicting Video ID in music video classification problems. ConvAE's LS and TPHM are input to DNN at separate instances to examine optimal features.

Finally, the output of the softmax activation function was connected to the target feature (video ID) [56]. In detail, two variants of the input were used to train the models, one with the original topology preserving head-maps, treated as baseline (phase D of Figure 1), and the other with the latent space of the ConvAE activated from the same head-maps. The rationale of the baseline was to demonstrate the utility of the latent space in the music video classification.

Initially, input data were shuffled to reduce variance and to make sure train/validation/test sets were representative of the entire data distribution. A holdout validation approach partitioned the shuffled topographic maps and related activated latent space inputs into the train, validation, and test sets with 75%, 12.5%, and 12.5%, respectively. The DNN used Adam as an optimization algorithm where the optimal learning rate was determined by hyperparameter tuning. The DNN was trained using an early stopping mechanism with a patience value set to 10. The training was halted if the validation loss did not show any improvement (decrease) for more than 10 epochs. Accuracy and F1-score were used as evaluation metrics to assess the performance of the DNN in fitting the 40 video categories. Accuracy is the fraction of the correct predictions for the data.

$$\text{Accuracy} = \frac{\text{Right predictions}}{\text{Total predictions}} \tag{9}$$

F1-score is the harmonic mean of the precision and recall. The value of a perfect F1-score is 1.

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{10}$$

3.6. ConvAE and DNN Hyperparameter Tuning

Hyperparameter tuning was carried out separately for both the ConvAE and DNN to optimize model performance. In detail, person-specific ConvAE and DNN were trained on

three random participants and following the pipeline of Figure 1. Several hyperparameters were manipulated, and the model trained highlighted the best configuration:

1. Three convolutional layers, each in both encoder and decoder, lead to the best reconstruction capacity of the ConvAE. No significant improvement in model performance was observed for more than three convolutional layers. Thus the network was not expanded. The Learning Rate (LR) scheduler gave 3×10^{-4} for the Adam optimizer. The optimal batch size was found to be 32. For encoders, performance was optimal when the number of kernels was doubled at each convolutional layer while image dimensions were halved. Symmetrically, the number of kernels was halved for decoders while dimensions doubled until the output layer, where the image size equaled the original input.
2. For DNN, five dense layers gave the optimal performance. The LR scheduler gave 3×10^{-4} for Adam optimizer. The optimal batch size was found to be 32. The Kullback–Leibler (KL) divergence outperformed other loss metrics including categorical cross-entropy for multiclass classification with softmax activation in the output layer with the one-hot encoded target variable (video ID). KL divergence is given by:

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) \log \frac{p(x_i)}{q(x_i)} \tag{11}$$

The problem of overfitting was handled with more data generation, *L2* regularization, and dropout regularization.

1. As mentioned previously, more data were generated by overlapping EEG signal windows using a window shift of 125 ms, 250 ms, and 500 ms.
2. L2 regularization was added to the convolutional layers of Autoencoder with the regularization factor tuned to 0.01.
3. Dropout regularization was introduced in DNN with a rate of 0.1.
4. Early stopping monitored training and validation epochs in both ConvAE and DNN. Model training was stopped when no significant decrease was found in the validation loss over 10 epochs.

3.7. Implementation Details

The initial training for a single participant was conducted in Google Colab with TensorFlow, Keras, and machine learning libraries, using automatically allocated GPUs. The experiment consisted of training 48 ConvAE models for window length, window shift, and latent space configurations to compute reconstruction metrics for each participant (4WL $\times$ 3WS $\times$ 4LS = 48 models). It also included 60 DNN models (48 + 12 using the original topographic maps). Thus a total of 108 models for each participant were built. Each model took, on average, between 45 and 55 min to train with Google Colab.

Training 108 models would have required between 4860 and 5940 min (81 to 99 h, which is 3.4 to 4.1 days) for each participant. Consequently, 108 to 132 days were required to run the entire experiment, with no interruptions and no accounting of potential technical problems. Thus, it was decided to move to a dedicated server for training the remaining participants. The server was an IBM machine with four Tesla P100 GPUs and 160 CPUs with a Linux kernel 167-Ubuntu SMP version with a ppc64le architecture.

3.8. Statistics

The DNN's evaluation metrics accuracy and F1-score were used to assess the utility of the various latent space dimensions by comparing before and after training each ConvAE on the original topographic maps for every window length and window shift configuration.

The one-sided Wilcoxon signed-rank test (with $\alpha = 0.05$) for equality of medians on the paired samples (same subject) was used to test the null hypothesis: the utility values for latent space parameters were less than or equal to the utility values for topographic

maps against the alternate hypothesis: the utility values for latent space parameters were greater than utility values for topographic maps.

4. Results

4.1. Convolutional Autoencoders (ConvAE)

Table 2 presents the results associated with the reconstruction metrics of the ConvAE coupled with the box-plots in Figure 4, and the 3D plots in Figure 5. As it is possible to notice, generally, all models achieved SSIM scores close to 1, which indicated that the trained models were able to reconstruct topographic head-maps structurally very similar to the original ones. The best-performing models were built with a window shift of 125 ms. Small window shifts generated more topographic maps, which increased the training data for the model to learn from, thereby improving the model performance by minimizing the training error (Figure 4). Intuitively, the latent space of (16,16,5) of the ConvAEs leads to models with a better SSIM score than its smaller counterparts of (4,4,5) and (2,2,5). In (4,4,5) and (2,2,5), the feature size was further reduced with convolution and pooling. The larger the latent space and the smaller the window shift, the higher the SSIM score. This pattern was observed in every window length of 0.5 s, 1 s, 1.5 s, and 2 s. However, the EEG window length does not affect the SSIM scores, as similar results were produced across each window length. Furthermore, the latent space of (8,8,5) gave promising results close to that of (16,16,5) for a 125 ms shift, signaling the potential of obtaining a similar reconstruction capacity with an additional reduction in the size of the input topographic maps.

Table 2. Reconstruction scores from Structural Similarity Index Measure (SSIM), Mean Squared Error (MSE), Normalized Root Mean Squared Error (NRMSE), and Peak Signal-to-Noise Ratio (PSNR) for 48 models obtained with every possible window length (WL), window shift (WS), and latent space (LS) combinations.

Model Configuration			Reconstruction Metric Scores			
WL (s)	**WS (ms)**	**LS**	**SSIM**	**MSE**	**NRMSE**	**PSNR**
0.5	125	(2,2,5)	0.9985	0.0000069	0.0361	55.75
0.5	125	(4,4,5)	0.9992	0.0000032	0.0279	58.14
0.5	125	(8,8,5)	0.9996	0.0000015	0.0206	60.78
0.5	125	(16,16,5)	0.9997	0.0000009	0.0158	63.32
0.5	250	(2,2,5)	0.9980	0.0000095	0.0461	53.56
0.5	250	(4,4,5)	0.9990	0.0000044	0.0345	56.27
0.5	250	(8,8,5)	0.9991	0.0000039	0.0312	57.48
0.5	250	(16,16,5)	0.9994	0.0000017	0.0220	60.60
0.5	500	(2,2,5)	0.9975	0.0000115	0.0510	52.89
0.5	500	(4,4,5)	0.9987	0.0000049	0.0369	55.42
0.5	500	(8,8,5)	0.9992	0.0000034	0.0303	57.25
0.5	500	(16,16,5)	0.9993	0.0000028	0.0257	59.20
1	125	(2,2,5)	0.9980	0.0000104	0.0384	54.27
1	125	(4,4,5)	0.9993	0.0000033	0.0251	57.98
1	125	(8,8,5)	0.9997	0.0000013	0.0169	61.23
1	125	(16,16,5)	0.9997	0.0000016	0.0182	61.39
1	250	(2,2,5)	0.9976	0.0000129	0.0442	52.97
1	250	(4,4,5)	0.9986	0.0000064	0.0360	55.35
1	250	(8,8,5)	0.9993	0.0000042	0.0276	57.17
1	250	(16,16,5)	0.9996	0.0000022	0.0211	59.36
1	500	(2,2,5)	0.9971	0.0000152	0.0514	51.74
1	500	(4,4,5)	0.9985	0.0000075	0.0391	53.92
1	500	(8,8,5)	0.9991	0.0000043	0.0308	56.31
1	500	(16,16,5)	0.9989	0.0000067	0.0326	56.47
1.5	125	(2,2,5)	0.9978	0.0000128	0.0385	53.36
1.5	125	(4,4,5)	0.9992	0.0000035	0.0241	57.09

Table 2. *Cont.*

Model Configuration			Reconstruction Metric Scores			
WL (s)	**WS (ms)**	**LS**	**SSIM**	**MSE**	**NRMSE**	**PSNR**
1.5	125	(8,8,5)	0.9996	0.0000017	0.0169	60.48
1.5	125	(16,16,5)	0.9997	0.0000013	0.0149	61.67
1.5	250	(2,2,5)	0.9972	0.0000163	0.0455	52.01
1.5	250	(4,4,5)	0.9987	0.0000065	0.0335	54.38
1.5	250	(8,8,5)	0.9994	0.0000032	0.0228	57.99
1.5	250	(16,16,5)	0.9995	0.0000022	0.0200	59.28
1.5	500	(2,2,5)	0.9963	0.0000193	0.0542	50.66
1.5	500	(4,4,5)	0.9980	0.0000100	0.0412	52.69
1.5	500	(8,8,5)	0.9991	0.0000042	0.0276	56.16
1.5	500	(16,16,5)	0.9992	0.0000038	0.0257	56.95
2	125	(2,2,5)	0.9984	0.0000085	0.0357	55.16
2	125	(4,4,5)	0.9992	0.0000033	0.0251	58.27
2	125	(8,8,5)	0.9997	0.0000014	0.0164	61.59
2	125	(16,16,5)	0.9998	0.0000010	0.0146	62.57
2	250	(2,2,5)	0.9979	0.0000101	0.0427	53.52
2	250	(4,4,5)	0.9988	0.0000047	0.0321	56.07
2	250	(8,8,5)	0.9994	0.0000023	0.0234	58.69
2	250	(16,16,5)	0.9991	0.0000037	0.0241	59.60
2	500	(2,2,5)	0.9973	0.0000127	0.0511	51.89
2	500	(4,4,5)	0.9980	0.0000078	0.0415	53.86
2	500	(8,8,5)	0.9987	0.0000051	0.0330	55.83
2	500	(16,16,5)	0.9990	0.0000033	0.0265	57.98

Similarly, the MSE results indicate that the larger the latent space, the smaller the reconstruction error, with (16,16,5) being the best-performing latent space. The latent space of (8,8,5) gives similar results to that of (16,16,5) for 125 ms shift (Figures 4 and 5). The window shift has a more evident trend and shares a positive correlation with MSE: the smaller the shift, the smaller the MSE. Window length again seems not to have any influence on the MSE scores. The results associated with the NRMSE confirmed the observation made with the MSE scores. In other words, the larger the latent space and smaller the window shift, the lower the NRMSE score, with window lengths not leading to observable effects on the NRMSE score. Finally, the results associated with the PSNR confirmed that the larger the latent space dimension, the higher the peak signal-to-noise ratio of the reconstructed topographic head-maps, with smaller window shifts better than larger shifts and window size not influencing the PSNR score.

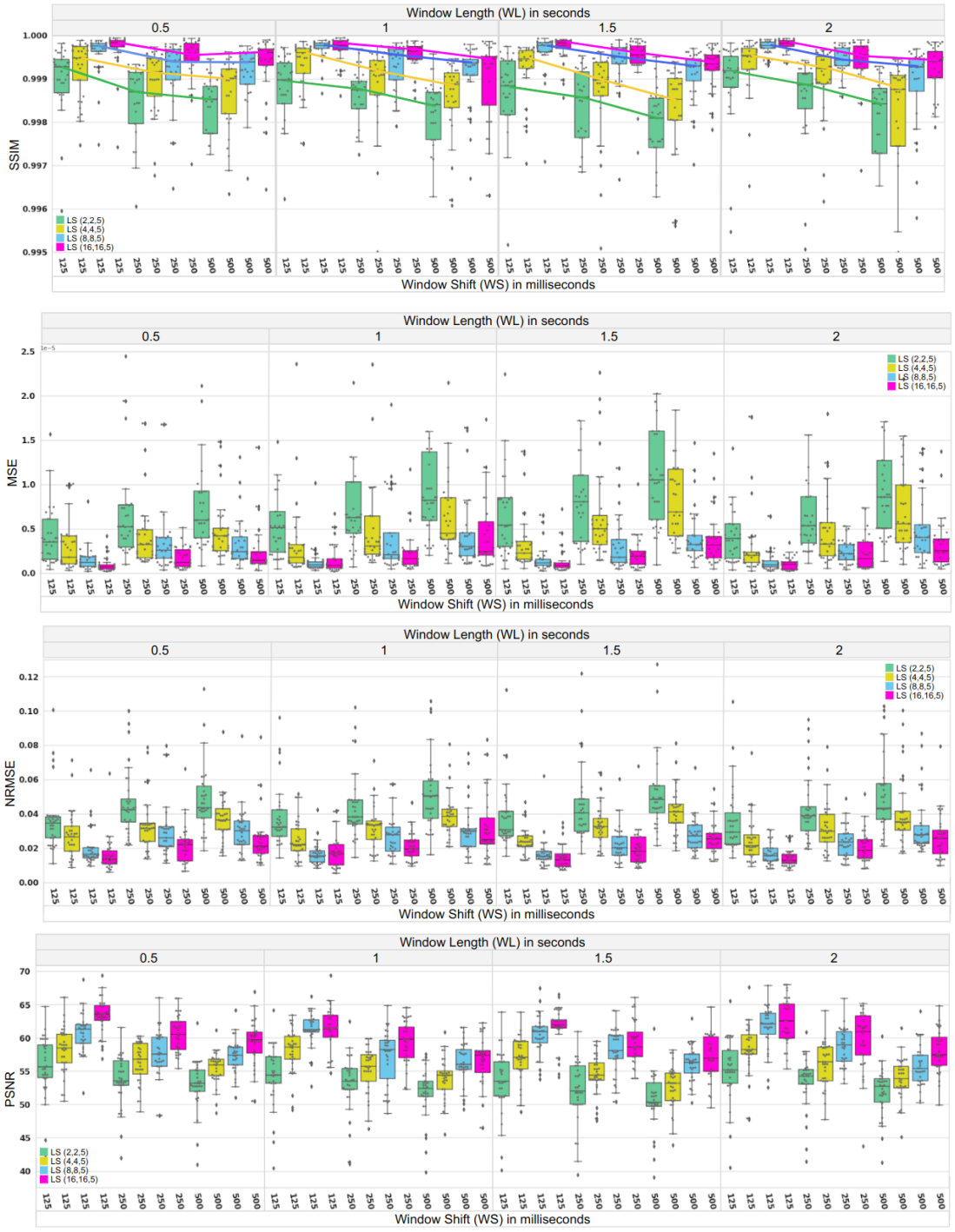

Figure 4. Reconstruction ability, as measured by Structural Similarity Index (SSIM), Mean Squared Error (MSE), Normalised Root Mean Squared Error (NRMSE), and Peak Signal-to-Noise Ratio (PSNR) for 48 different ConvAE models with varying window length (WL), window shift (WS), and latent space (LS).

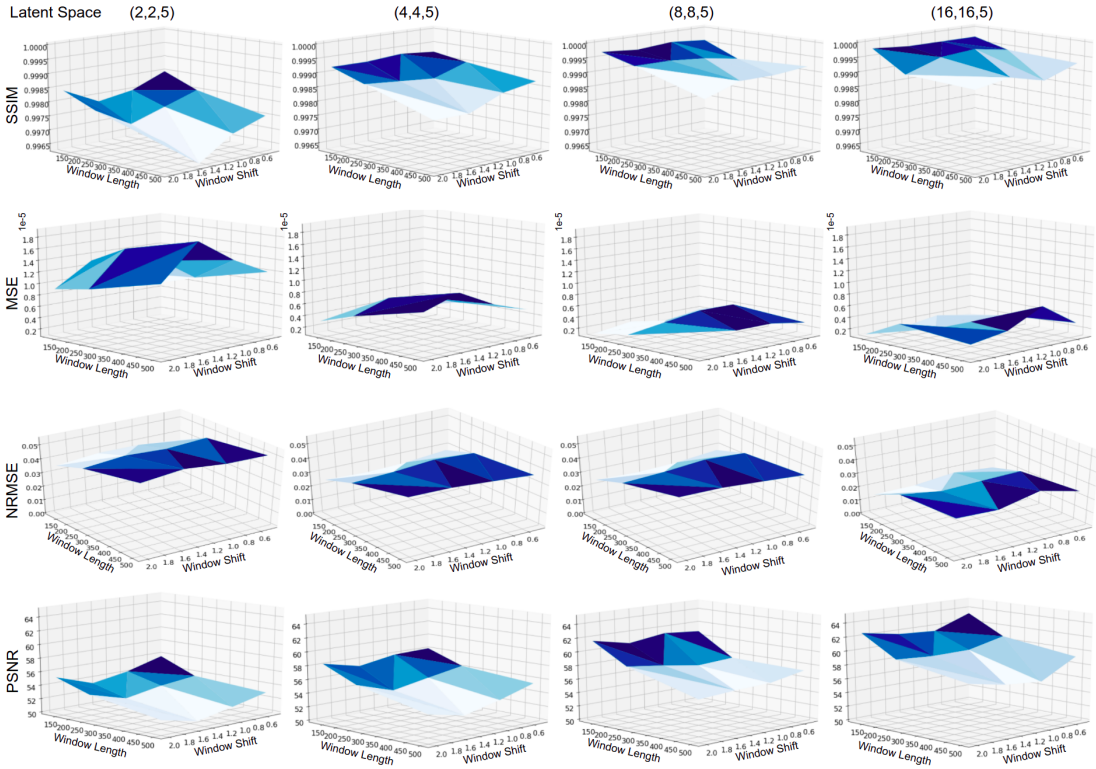

Figure 5. 3D plot depicting mean reconstruction scores (SSIM, MSE, NRMSE, and PSNR) for 48 different ConvAE models with varying window length (WL), window shift (WS), and latent space (LS).

4.2. Dense Neural Network (DNN)

The reconstruction metrics confirmed an intuitive trend whereby the less information cut and the more training instances yield higher reconstruction capacity. However, the results in this section aim to demonstrate whether such an information cut and variation in training size affect the utility of the learned latent spaces of the various ConvAEs. Figure 6 compares the utility scores for DNN models in the form of a box-plot, while Figure 7 presents them as 3D plots. Every latent space of the models leads to better predictive accuracy than those models trained with the original (full size) topographic maps (Figure 6) except for (2,2,5) latent space with 500 ms window shift in 0.5 s and 1 s window length case. The latent space of (2,2,5) with a 500 ms window shift leads to the creation of models with accuracy in predicting video categories comparable to the original, full-size topographic maps of (32,32,5). In general, the larger the latent space, the better the predictive accuracy of the DNN models, with the accuracies of models built with latent space of (8,8,5) and (16,16,5) nearly equal (Table 3, Figures 8 and 9).

The 125 ms window shift leads to models with the best accuracies, thereby adhering to the previously established results that the smaller the window shift better the accuracy. Finally, a larger window length improved the predictive accuracy of video categories with a significant improvement observed from 0.5 s to 1 s window length, indicating that the minimum length of the EEG windows has to be at least 1 second, with the current sampling rate (128 Hz) to achieve a good performance.

Observation for F1-scores was in line with the accuracy scores. Every latent space leads to creating models with better F1-scores than those built with full-size original input topographic maps except for (2,2,5) latent space with 500 ms window shift in 0.5 s and 1 s window length case. The latent space of (2,2,5) with a 500 ms window shift leads to the creation of models with an F1-score in predicting video categories comparable to the original, full-size topographic maps of (32,32,5). The latent spaces of (8,8,5) and (16,16,5) lead to the formation of models with the highest F1-scores. Similarly, small window shifts lead to better F1-scores. The larger the window length, the better the F1-score, with 1 s

being the minimum length of the EEG windows, with the current sampling rate (128 Hz) to achieve good performance.

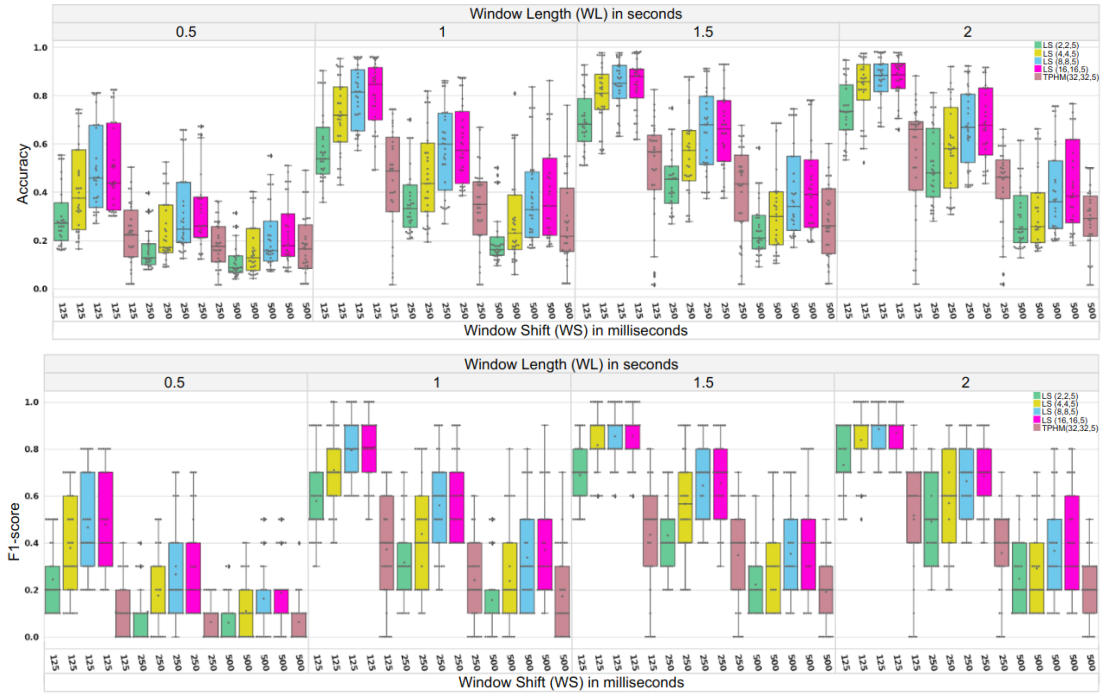

Figure 6. Utility (U) scores (Accuracy, F1-score) for 60 different DNN models with varying window length (WL), window shift (WS), and two kinds of input viz latent space (LS) and Topology Preserved Head-Maps (TPHM).

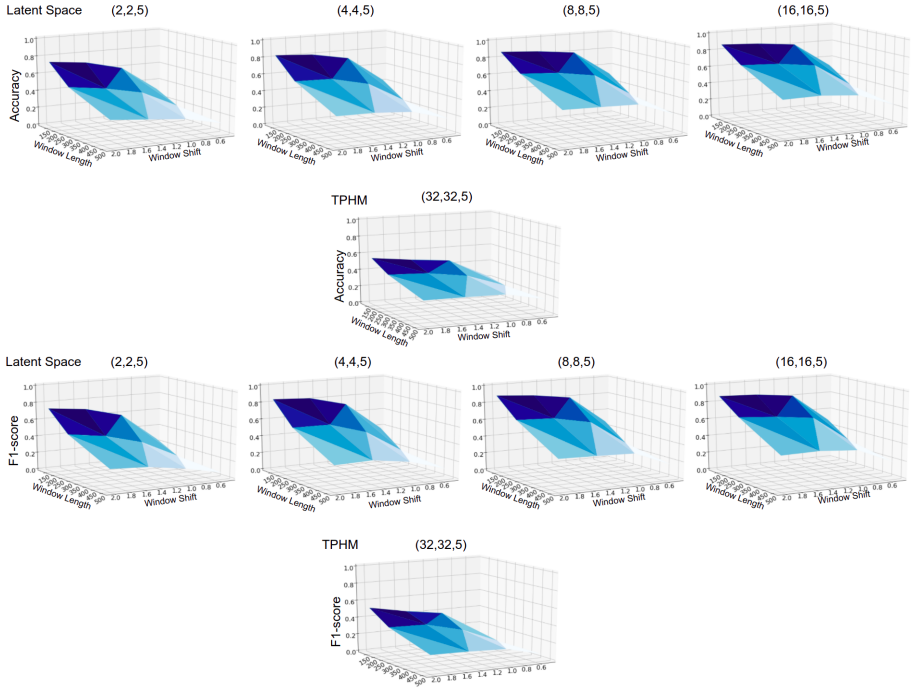

Figure 7. 3D plot depicting mean Utility (U) scores (Accuracy, F1-score) for 60 different DNN models with varying window length (WL), window shift (WS), and two kinds of input viz latent space (LS) and Topology Preserved Head-Maps (TPHM).

Table 3. Average accuracies of the classification models for all the participants trained with latent spaces (LS) and the original topographic maps (TPHM), grouped by window length (WL) for the optimal window shift of 125 ms.

WL	LS	Avg. Acc.	TPHM	Avg. Acc.
0.5	(2,2,5) (4,4,5) (8,8,5) (16,16,5)	29.9% 41.5% 49.9% 50.6%	(32,32,5)	22.4%
1	(2,2,5) (4,4,5) (8,8,5) (16,16,5)	59.0% 72.0% 79.1% 79.9%	(32,32,5)	44.8%
1.5	(2,2,5) (4,4,5) (8,8,5) (16,16,5)	69.1% 79.8% 83.9% 84.2%	(32,32,5)	49.0%
2	(2,2,5) (4,4,5) (8,8,5) (16,16,5)	73.3% 82.4% 86.7% 86.6%	(32,32,5)	54.3%

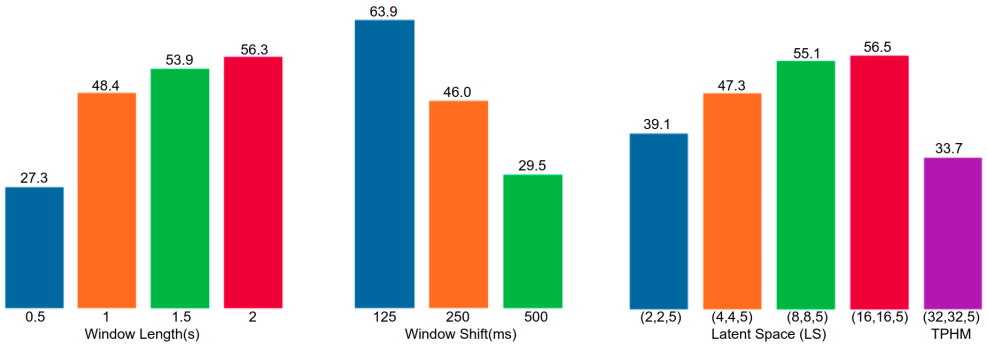

Figure 8. Aggregate mean accuracy percentage of the 60 models on window length (WL), window shift (WS), and latent space (LS).

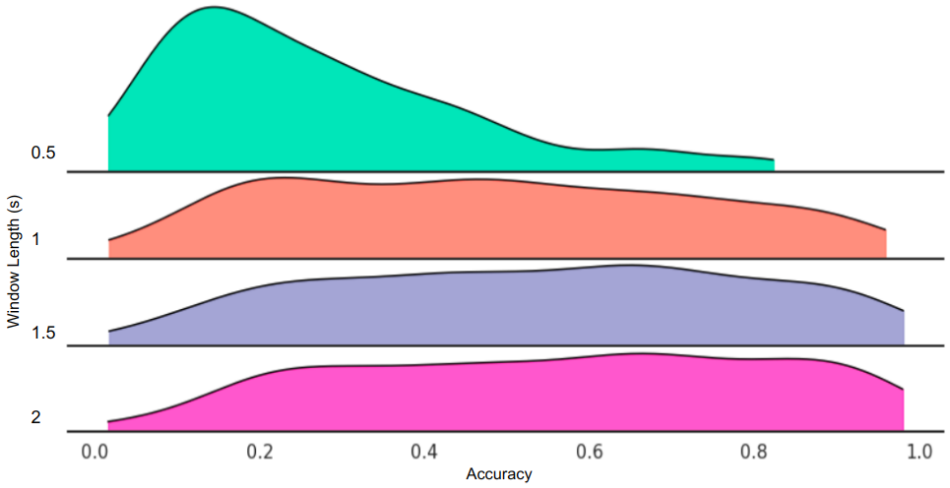

Figure 9. *Cont.*

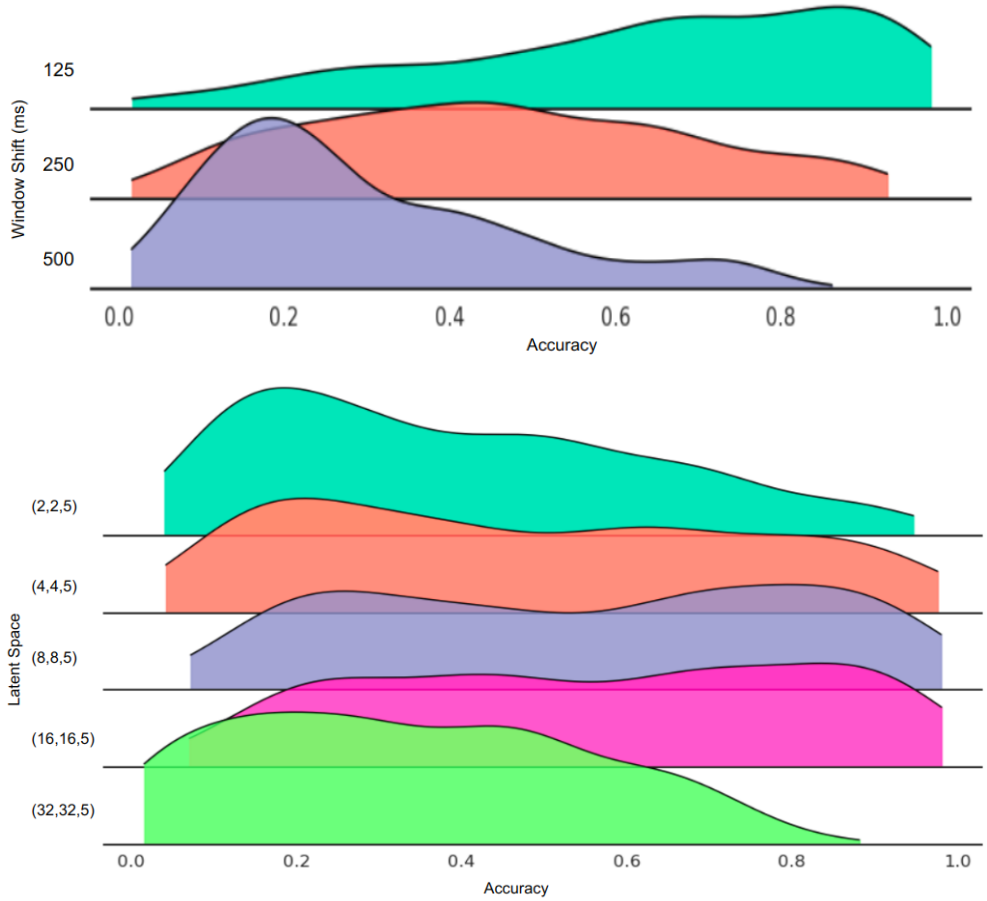

Figure 9. Accuracy distribution of the classification models for all the participants across window length, window shift, and latent space.

4.3. Statistical Inferences

Table 4 presents the p-values associated with the pair-wise comparison of the distributions of the predictive accuracies and the F1-scores of the DNN models trained respectively with latent spaces and the original, full-size topographic maps using the Wilcoxon signed-rank test. The results were statistically significant for 125 ms window shift irrespective of window length and latent space. As we increased the window shift to 250 ms and 500 ms, few comparisons with latent spaces of (2,2,5) and (4,4,5) became non-significant. The non-significant p-value suggests that, despite achieving good reconstruction scores in reconstructing topographic EEG maps, the DNN models trained with latent space of (2,2,5) and (4,4,5) dimensions were ineffective.

Table 4. Comparison of utility scores between two groups (topology preserved head-maps and each latent space) for a population size of 32 in a paired one-sided Wilcoxon signed-rank test with $\alpha = 0.05$ (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

LS	TPHM	*p*-Value for Accuracy			*p*-Value for F1-Score		
		WS 125 ms	WS 250 ms	WS 500 ms	WS 125 ms	WS 250 ms	WS 500 ms
				WL 0.5 s			
(2,2,5)	(32,32,5)	0.039 *	0.839	0.999	2.637×10^{-5} ***	0.064	0.683
(4,4,5)	(32,32,5)	5.076×10^{-6} ***	0.017 *	0.982	1.626×10^{-6} ***	9.219×10^{-5} ***	0.051
(8,8,5)	(32,32,5)	4.787×10^{-7} ***	1.287×10^{-05} ***	0.102	5.406×10^{-7} ***	2.290×10^{-6} ***	0.0001 ***
(16,16,5)	(32,32,5)	5.264×10^{-7} ***	2.103×10^{-6} ***	0.009 **	6.301×10^{-7} ***	6.863×10^{-7} ***	2.355×10^{-5} ***

Table 4. *Cont.*

LS	TPHM	*p*-Value for Accuracy			*p*-Value for F1-Score		
		WS 125 ms	WS 250 ms	WS 500 ms	WS 125 ms	WS 250 ms	WS 500 ms
				WL 1 s			
(2,2,5)	(32,32,5)	0.001 **	0.492	0.994	0.0001 ***	0.071	0.752
(4,4,5)	(32,32,5)	5.533×10^{-6} ***	0.004 **	0.537	2.614×10^{-6} ***	0.0003 ***	0.035 *
(8,8,5)	(32,32,5)	5.264×10^{-7} ***	3.281×10^{-6} ***	0.0005 ***	1.246×10^{-6} ***	4.817×10^{-6} ***	2.096×10^{-5} ***
(16,16,5)	(32,32,5)	4.352×10^{-7} ***	4.787×10^{-7} ***	4.120×10^{-5} ***	8.376×10^{-7} ***	1.226×10^{-6} ***	2.591×10^{-6} ***
				WL 1.5 s			
(2,2,5)	(32,32,5)	0.0004 ***	0.239	0.969	4.675×10^{-5} ***	0.112	0.322
(4,4,5)	(32,32,5)	1.112×10^{-6} ***	0.0002 ***	0.281	8.254×10^{-7} ***	0.0001 ***	0.003 **
(8,8,5)	(32,32,5)	3.954×10^{-7} ***	5.533×10^{-6} ***	0.0003 ***	8.281×10^{-7} ***	3.642×10^{-6} ***	2.555×10^{5} ***
(16,16,5)	(32,32,5)	4.352×10^{-7} ***	1.219×10^{-6} ***	3.392×10^{-5} ***	8.445×10^{-7} ***	1.797×10^{-6} ***	1.929×10^{-5} ***
				WL 2 s			
(2,2,5)	(32,32,5)	0.001 **	0.043 *	0.631	0.0002 ***	0.010 *	0.050
(4,4,5)	(32,32,5)	3.582×10^{-6} ***	0.001 **	0.139	4.083×10^{-6} ***	0.0006 ***	0.008 **
(8,8,5)	(32,32,5)	6.443×10^{-7} ***	1.756×10^{-6} ***	0.001 **	1.211×10^{-6} ***	1.843×10^{-6} ***	0.0001 ***
(16,16,5)	(32,32,5)	3.954×10^{-7} ***	5.264×10^{-7} ***	0.0001 ***	1.241×10^{-6} ***	1.226×10^{-6} ***	2.363×10^{-5} ***

5. Discussion

In general, the results associated with the person-specific ConvAE highlighted that:

(I) the larger the latent space, the higher the reconstruction ability,
(II) the smaller the window shift, the higher the reconstruction ability,
(III) window length did not have an important role, and it did not influence the reconstruction ability,
(IV) on average, the utility of all the latent spaces learned in each ConvAE outperformed that associated with the original topographic maps;
(V) the best utility of the latent space is when the input is of shape (8,8,5) with window shift 125 ms and with a window length of at least 1 s.

For finding (I), the fact that the reconstruction ability is higher given larger latent spaces is intuitively explained by the amount of information each latent space holds. (16,16,5) had 1280 data points, (8,8,5), (4,4,5), and (2,2,5) had 320, 80, and 20 points respectively. The larger latent space encompasses more meaningful internal representations of the externally observed topographic maps—the inputs of ConvAE.

For finding (II), the smaller the window shift, the higher the reconstruction ability is explained by the variability across the generated topographic maps. For example, for a 1 s window length, 473 topographic maps were generated for a shift of 125 ms, whereas for 250 ms and 500 ms window shifts, 237 and 119 topographic maps were generated, respectively. As encoded in a topographic map, more variation in cerebral activation exists for smaller shifts. In other words, different brain dynamics can be represented and used to train ConvAE with varying topographic maps.

For finding (III), the fact that the length of the window does not affect the reconstruction capacity of the ConvAE is explained reasonably by the number of topographic maps used for training each person-specific ConvAE. For example, by taking 125 ms as the fixed value for a window shift, 477 topographic maps were produced with a 0.5 s window length. The topographic maps generated for 1 s, 1.5 s, and 2 s window lengths were 473, 469, and 465, respectively. The cardinality of the training sets of the generated topographic maps differs minimally across window lengths and their variability. Therefore, it can be argued that reconstruction capacity across the different window lengths could be attributed to the amount of training data. To verify this was untrue, all the person-specific ConvAE's were retrained by augmenting each training set by adding Gaussian noise. In detail, for each training tensor of (32,32,5) in the training set, one, two, and three augmented tensors were

generated by adding Gaussian noise (mean 0, std 1) [57]. The cardinality of each original training set was increased by augmenting the data once, twice, and thrice respectively. The results of this additional augmentation phase were consistent with the previous results, whereby the manipulation of the window length did not improve the reconstruction ability of each person-specific ConvAE.

For finding (IV), every latent space of the ConvAE outperformed the original topographic maps when trained with a DNN to fit the selected classification task (video categorization), clearly indicating its utility. The latent space likely contains an increasing number of relevant representations of the original topographic maps, and they have likely discarded their inherent noise and artifacts. Both utility metrics, namely accuracy and F1-score, were in unison with the established results: larger window length, smaller window shift, and bigger latent space delivered a better utility score. Furthermore, the analyses concerning the window lengths align with the results documented in [58] that uses a variational autoencoder. Here, intuitively, the larger the window length, the better its impact on the utility. However, it is beneficial to know the minimum amount of information in the latent space of the Autoencoder that reconstructs its inputs successfully. Specifically, our study finds the minimal window length, and the shift among windows required to train a convolutional autoencoder.

For finding (V), the combination of latent space of shape (8,8,5) with 125 ms shift and a minimal window length of 1 s can be justified by the following observations. Firstly, (8,8,5) means a reduction of the dimensionality of the original topographic maps by 75%, which is already a significant cut in information and, on average, gives nearly the same outcome as obtained with (16,16,5). The latent space (8,8,5) is not only a necessary dimension to capture the relevant features from the original data but also not larger for accommodating redundant features. Reducing the latent space dimension further starts deteriorating the utility faster. Secondly, as mentioned before, 125 ms allows more training instances to create a substantial variability in cerebral activation that is encoded in additional topographic maps. Intuitively, this variability in the input might be transitively expected in the activation of the latent space of trained ConvAE, thus positively impacting the discrimination of the video categories. Thirdly, 1 s is the minimal length for a window to contain relevant information for discriminating video, presumably because of the FFT and the current sampling rate in EEG data. In fact, with a sampling rate of 128 Hz, the FFT of the input EEG signals can generate richer information in the frequency domain with at least a 1 s window compared to that obtained from a 0.5 s window (only 64 points). Consequently, the window length can likely affect the construction of precise topographic maps that better reflect cerebral activity in the frequency domain and the induction of robust latent spaces, maximizing their utility in categorizing videos.

Our research demonstrated that using a sliding window technique, spectral topographic head-maps can be formed from multichannel EEG signals, and ConvAE can be trained to extract relevant features, thus performing meaningful dimensionality reduction. The study illustrates the existence of an optimal window length (WL) of 1 s, at the mostly adopted 128 Hz sampling rate, an optimal 125 ms window shift (WS), and an optimal latent space (LS) of 25% the original size, that can maximize the mean reconstruction capacity of the spectral topographic head-maps via trained ConvAE models and that has maximal utility in a classification task.

Our study contributes to the body of knowledge with an architectural pipeline for eliminating redundant EEG data while preserving relevant EEG features with deep autoencoders and establishing its limits and utility that can be used in ecological settings. Here ecological settings mean naturalistic tasks performed by users like those executed in the study that leads to the development of the DEAP dataset [51].

6. Conclusions

Electroencephalographic (EEG) signals can be analyzed in various domains, including time, space, and frequency. However, noise and artifacts exist in these signals, including

eye and muscle movements, to mention a few, adding difficulties to their analysis. For this reason, techniques such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are used to detect and remove artifacts requiring human intervention. Autoencoders, an unsupervised method that does not require human labeling, have automatized artifact detection and removal by representing inputs in a lower dimensional latent space. However, little research is devoted to understanding the minimum dimension of such latent space that allows meaningful input reconstruction.

An empirical experiment was conducted to investigate the optimal latent space dimension with continuous EEG data gathered from 32 subjects while watching 40 music video excerpts chosen to evoke different emotions in participants, thus exerting different mental states. Person-specific ConvAE architectures were constructed by manipulating the size of its latent space. A sliding window technique has been employed by segmenting continuous EEG signals into windows of varying sizes and employing an overlapping strategy. Five topographic head-maps were formed for every window in the frequency domain, one for each EEG band (delta, theta, alpha, beta, and gamma). The latent space of autoencoders was assessed according to the topographic maps' reconstruction capacity and its utility in classifying the 40 videos. Findings suggest that the minimal latent space dimension is 25% of the size of the input topographic maps for achieving maximum reconstruction capacity and maximizing classification accuracy. In detail, this was achieved with a window length of at least 1 s and a shift of 125 ms with the sampling rate of 128 Hz.

Our study contributes to the body of knowledge with an architectural pipeline for eliminating redundant EEG data while preserving relevant EEG features with deep autoencoders and establishing its limits and utility that can be used in ecological settings. Future work will include the extension of this study to further strengthen its findings. For example, larger time windows can be tested to understand further their impact on the reconstruction capacity of autoencoders and the utility of the derived latent spaces. Similarly, smaller shifts, and larger topographic head-maps, built with fewer or more electrodes can be tested, and the current experiment can be replicated by employing different sampling rates. The architectural pipeline can be applied to individual and different combinations of the EEG bands (delta, theta, alpha, beta, gamma) to understand their impact on reconstruction capacity and predictive utilities. Notions of explainability and techniques from explainable artificial intelligence (XAI) can be employed [59] to analyze derived latent space with visual explanations [60], including salient masks or using techniques such as the layer-wise relevance propagation, which will provide scholars with a richer analysis tool and help improve the design of the autoencoder architectures.

Author Contributions: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, A.V.C. and L.L.; Supervision, L.L.; Validation, A.V.C. and L.L.; Visualization, A.V.C.; Writing—original draft, A.V.C.; Writing—review & editing, A.V.C. and L.L.; All authors have read and agreed to the published version of the manuscript.

Funding: This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Data Availability Statement: DEAP dataset: a dataset for emotion analysis using eeg, physiological and video signals https://www.eecs.qmul.ac.uk/mmv/datasets/deap/ (accessed on 10 January 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EEG	Electroencephalogram
CNN	Convolution Neural Network
PCA	Principal Component Analysis
ICA	Independent Component Analysis
WL	Window Length
LS	Latent Space
WS	Window Shift
TPHM	Topology Preserved Head Maps
ConvAE	Convolutional Autoencoder
SSIM	Structural Similarity Index Measure
MSE	Mean Square Error
NRMSE	Normalized Root-Mean-Square Error
PSNR	Peak Signal-to-Noise Ratio
DNN	Dense Neural Network
LDA	Linear Discriminant Analysis
CSP	Common Spatial Pattern
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
RF	Random Forest
GMM	Gaussian Mixture Model
DEAP	A dataset for emotion analysis using eeg, physiological and video signals
EOG	Electrooculogram
FFT	Fast Fourier Transform

References

1. Mars, R.B.; Sotiropoulos, S.N.; Passingham, R.E.; Sallet, J.; Verhagen, L.; Khrapitchev, A.A.; Sibson, N.; Jbabdi, S. Whole brain comparative anatomy using connectivity blueprints. *eLife* **2018**, *7*, e35237. [CrossRef] [PubMed]
2. Cohen, M.X. *Analyzing Neural Time Series Data: Theory and Practice*; MIT Press: Cambridge, MA, USA, 2014.
3. Alçin, Ö.F.; Siuly, S.; Bajaj, V.; Guo, Y.; Şengu, A.; Zhang, Y. Multi-Category EEG Signal Classification Developing Time-Frequency Texture Features Based Fisher Vector Encoding Method. *Neurocomputing* **2016**, *218*, 251–258. [CrossRef]
4. Stober, S.; Sternin, A.; Owen, A.M.; Grahn, J.A. Deep Feature Learning for EEG Recordings. *arXiv* **2015**, arxiv:1511.04306.
5. Férat, V.; Seeber, M.; Michel, C.M.; Ros, T. Beyond broadband: Towards a spectral decomposition of electroencephalography microstates. *Hum. Brain Mapp.* **2022**, *43*, 3047–3061. [CrossRef]
6. Abdeljaber, O.; Avcı, O.; Kiranyaz, M.S.; Boashash, B.; Sodano, H.A.; Inman, D.J. 1-D CNNs for structural damage detection: Verification on a structural health monitoring benchmark data. *Neurocomputing* **2018**, *275*, 1308–1317. [CrossRef]
7. Abdi, H.; Williams, L.J. Principal component analysis. *Wires Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
8. Bro, R.; Smilde, A.K. Principal component analysis. *Anal. Methods* **2014**, *6*, 2812–2831. [CrossRef]
9. Acharya, U.R.; Sree, S.V.; Swapna, G.; Martis, R.J.; Suri, J.S. Automated EEG analysis of epilepsy: A review. *Knowl. Based Syst.* **2013**, *45*, 147–165. [CrossRef]
10. Oosugi, N.; Kitajo, K.; Hasegawa, N.; Nagasaka, Y.; Okanoya, K.; Fujii, N. A New Method for Quantifying the Performance of EEG Blind Source Separation Algorithms by Referencing a Simultaneously Recorded ECoG Signal. *Neural Netw.* **2017**, *93*, 1–6. [CrossRef]
11. Korats, G.; Cam, S.L.; Ranta, R.; Hamid, M.R. Applying ICA in EEG: Choice of the Window Length and of the Decorrelation Method. In Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies—BIOSTEC, Vilamoura, Portugal, 1–4 February 2012.
12. Brunner, C.; Naeem, M.; Leeb, R.; Graimann, B.; Pfurtscheller, G. Spatial Filtering and Selection of Optimized Components in Four Class Motor Imagery EEG Data Using Independent Components Analysis. *Pattern Recogn. Lett.* **2007**, *28*, 957–964. [CrossRef]
13. Xing, X.; Li, Z.; Xu, T.; Shu, L.; Hu, B.; Xu, X. SAE+LSTM: A New Framework for Emotion Recognition From Multi-Channel EEG. *Front. Neurorobot.* **2019**, *13*, 37. [CrossRef] [PubMed]
14. Zhang, S.; You, B.; Lang, X.; Zhou, Y.; An, F.; Dai, Y.; Liu, Y. Efficient Rejection of Artifacts for Short-Term Few-Channel EEG Based on Fast Adaptive Multidimensional Sub-Bands Blind Source Separation. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–16. [CrossRef]
15. Hsu, S.H.; Mullen, T.; Jung, T.P.; Cauwenberghs, G. Real-Time Adaptive EEG Source Separation Using Online Recursive Independent Component Analysis. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2015**, *24*, 1. [CrossRef] [PubMed]

16. You, S.D.; Li, Y.C. Predicting Viewer's Preference for Music Videos Using EEG Dataset. In Proceedings of the 2020 IEEE International Conference on Consumer Electronics—Asia (ICCE-Asia), Seoul, Republic of Korea, 1–3 November 2020; pp. 1–2. [CrossRef]
17. Arabshahi, R.; Rouhani, M. A convolutional neural network and stacked autoencoders approach for motor imagery based brain-computer interface. In Proceedings of the 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 29–30 October 2020; pp. 295–300. [CrossRef]
18. Zhang, P.; Wang, X.; Zhang, W.; Chen, J. Learning Spatial–Spectral–Temporal EEG Features with Recurrent 3D Convolutional Neural Networks for Cross-Task Mental Workload Assessment. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2019**, *27*, 31–42. [CrossRef] [PubMed]
19. Yao, Y.; Plested, J; Gedeon, T. Deep Feature Learning and Visualization for EEG Recording Using Autoencoders. In Proceedings of the 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, 13–16 December 2018; Proceedings, Part VII.
20. Gaur, P.; Gupta, H.; Chowdhury, A.; McCreadie, K.; Pachori, R.B.; Wang, H. A sliding window common spatial pattern for enhancing motor imagery classification in EEG-BCI. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–9. [CrossRef]
21. Wilaiprasitporn, T.; Ditthapron, A.; Matchaparn, K.; Tongbuasirilai, T.; Banluesombatkul, N.; Chuangsuwanich, E. Affective EEG-based person identification using the deep learning approach. *IEEE Trans. Cogn. Dev. Syst.* **2019**, *12*, 486–496. [CrossRef]
22. Wang, X.; Wang, X.; Liu, W.; Chang, Z.; Kärkkäinen, T.J.; Cong, F. One dimensional convolutional neural networks for seizure onset detection using long-term scalp and intracranial EEG. *Neurocomputing* **2021**, *459*, 212–222. [CrossRef]
23. Huang, L.; Zhao, Y.; Zeng, Y.; Lin, Z. BHCR: RSVP target retrieval BCI framework coupling with CNN by a Bayesian method. *Neurocomputing* **2017**, *238*, 255–268. [CrossRef]
24. Qiu, Z.; Jin, J.; Lam, H.K.; Zhang, Y.; Wang, X.; Cichocki, A. Improved SFFS method for channel selection in motor imagery based BCI. *Neurocomputing* **2016**,*207*, 519–527. [CrossRef]
25. Sadatnejad, K.; Ghidary, S.S. Kernel learning over the manifold of symmetric positive definite matrices for dimensionality reduction in a BCI application. *Neurocomputing* **2016**, *179*, 152–160. [CrossRef]
26. Fei, Z.; Yang, E.; Li, D.D.U.; Butler, S.; Ijomah, W.; Li, X.; Zhou, H. Deep convolution network based emotion analysis towards mental health care. *Neurocomputing* **2020**, *388*, 212–227. [CrossRef]
27. Kurup, A.R.; Ajith, M.; Ramón, M.M. Semi-supervised facial expression recognition using reduced spatial features and Deep Belief Networks. *Neurocomputing* **2019**, *367*, 188–197. [CrossRef]
28. Xin Zhang, Y.; Chen, Y.; Gao, C. Deep unsupervised multi-modal fusion network for detecting driver distraction. *Neurocomputing* **2021**, *421*, 26–38. [CrossRef]
29. Yin, Z.; Zhao, M.; Zhang, W.; Wang, Y.; Wang, Y.; Zhang, J. Physiological-signal-based mental workload estimation via transfer dynamical autoencoders in a deep learning framework. *Neurocomputing* **2019**, *347*, 212–229. [CrossRef]
30. Ieracitano, C.; Mammone, N.; Bramanti, A.; Hussain, A.; Morabito, F.C. A Convolutional Neural Network approach for classification of dementia stages based on 2D-spectral representation of EEG recordings. *Neurocomputing* **2019**, *323*, 96–107. [CrossRef]
31. Su, R.; Liu, T.; Sun, C.; Jin, Q.; Jennane, R.; Wei, L. Fusing convolutional neural network features with hand-crafted features for osteoporosis diagnoses. *Neurocomputing* **2020**, *385*, 300–309. [CrossRef]
32. Chambon, S.; Galtier, M.N.; Arnal, P.J.; Wainrib, G.; Gramfort, A. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 758–769. [CrossRef]
33. Lee, S.B.; Kim, H.J.; Kim, H.; Jeong, J.H.; Lee, S.W.; Kim, D.J. Comparative analysis of features extracted from EEG spatial, spectral and temporal domains for binary and multiclass motor imagery classification. *Inf. Sci.* **2019**, *502*, 190–200. [CrossRef]
34. Li, Z.; Wang, J.; Jia, Z.; Lin, Y. Learning Space-Time-Frequency Representation with Two-Stream Attention Based 3D Network for Motor Imagery Classification. In Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM), Sorrento, Italy, 17–20 November 2020; pp. 1124–1129. [CrossRef]
35. Ang, K.K.; Chin, Z.Y.; Wang, C.; Guan, C.; Zhang, H. Filter Bank Common Spatial Pattern Algorithm on BCI Competition IV Datasets 2a and 2b. *Front. Neurosci.* **2012**, *6*. [CrossRef]
36. Subasi, A.; Gursoy, M.I. EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Syst. Appl.* **2010**, *37*, 8659–8666. [CrossRef]
37. Jirayucharoensak, S.; Pan-Ngum, S.; Israsena, P. EEG-Based Emotion Recognition Using Deep Learning Network with Principal Component Based Covariate Shift Adaptation. *Sci. World J.* **2014**, *2014*, 627892. [CrossRef] [PubMed]
38. Viola, F.C.; Debener, S.; Thorne, J.; Schneider, T.R. Using ICA for the analysis of multi-channel EEG data. In *Simultaneous EEG and fMRI: Recording, Analysis, and Application: Recording, Analysis, and Application*; Oxford Academic, New York, NY, USA, 2010; pp. 121–133.
39. Lemm, S.; Blankertz, B.; Curio, G.; Muller, K.R. Spatio-spectral filters for improving the classification of single trial EEG. *IEEE Trans. Biomed. Eng.* **2005**, *52*, 1541–1548. [CrossRef] [PubMed]
40. Wu, W.; Chen, Z.; Gao, X.; Li, Y.; Brown, E.N.; Gao, S. Probabilistic common spatial patterns for multichannel EEG analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 639–653. [CrossRef] [PubMed]
41. Qi, Y.; Luo, F.; Zhang, W.; Wang, Y.; Chang, J.; Woodward, D.; Chen, A.; Han, J. Sliding-window technique for the analysis of cerebral evoked potentials. **2003**, *35*, 231–235.

42. Alickovic, E.; Kevric, J.; Subasi, A. Performance evaluation of empirical mode decomposition, discrete wavelet transform, and wavelet packed decomposition for automated epileptic seizure detection and prediction. *Biomed. Signal Process. Control.* **2018**, *39*, 94–102. [CrossRef]
43. Atkinson, J.; Campos, D. Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers. *Expert Syst. Appl.* **2016**, *47*, 35–41. [CrossRef]
44. Edelman, B.; Baxter, B.; He, B. EEG source imaging enhances the decoding of complex right-hand motor imagery tasks. *Ire Trans. Med. Electron.* **2016**, *63*, 4–14. [CrossRef]
45. Faust, O.; Acharya, U.R.; Adeli, H.; Adeli, A. Wavelet-based EEG processing for computer-aided seizure detection and epilepsy diagnosis. *Seizure* **2015**, *26*, 56–64. [CrossRef]
46. Katsigiannis, S.; Ramzan, N. DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 98–107. [CrossRef]
47. Dargan, S.; Kumar, M.; Ayyagari, M.R.; Kumar, G. A survey of deep learning and its applications: a new paradigm to machine learning. *Arch. Comput. Methods Eng.* **2020**, *27*, 1071–1092. [CrossRef]
48. Zhang, L.; Tan, J.; Han, D.; Zhu, H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discov. Today* **2017**, *22*, 1680–1685. [CrossRef] [PubMed]
49. Bank, D.; Koenigstein, N.; Giryes, R. Autoencoders. *arXiv* **2020**, arXiv:2003.05991.
50. Li, J.; Struzik, Z.R.; Zhang, L.; Cichocki, A. Feature learning from incomplete EEG with denoising autoencoder. *arXiv* **2015**, arXiv:1410.0818.
51. Koelstra, S.; Muhl, C.; Soleymani, M.; Jong-Seok Lee.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Trans. Affect. Comput.* **2012**, *3*, 18–31. [CrossRef]
52. Ng, A. Sparse autoencoder. *CS294A Lecture Notes.* **2011**, *72*, 1–19.
53. Ahlawat, S.; Choudhary, A.; Nayyar, A.; Singh, S.; Yoon, B. Improved handwritten digit recognition using convolutional neural networks (CNN). *Sensors* **2020**, *20*, 3344. [CrossRef]
54. Sara, U.; Akter, M.; Uddin, M.S. Image quality assessment through FSIM, SSIM, MSE and PSNR—A comparative study. *J. Comput. Commun.* **2019**, *7*, 8–18. [CrossRef]
55. Daoud, H.; Bayoumi, M. Deep Learning Approach for Epileptic Focus Localization. *IEEE Trans. Biomed. Circuits Syst.* **2020**, *14*, 209–220. [CrossRef]
56. Abdelhameed, A.M.; Daoud, H.G.; Bayoumi, M. Epileptic Seizure Detection using Deep Convolutional Autoencoder. In Proceedings of the 2018 IEEE International Workshop on Signal Processing Systems (SiPS), Cape Town, South Africa, 21–24 October 2018; pp. 223–228. [CrossRef]
57. Hussain, Z.; Gimenez, F.; Yi, D.; Rubin, D. Differential data augmentation techniques for medical imaging classification tasks. In Proceedings of the AMIA Annual Symposium Proceedings, American Medical Informatics Association, Washington, DC, USA, 4 November 2017; Volume 2017, p. 979.
58. Ahmed, T.; Longo, L. Examining the Size of the Latent Space of Convolutional Variational Autoencoders Trained With Spectral Topographic Maps of EEG Frequency Bands. *IEEE Access* **2022**, *10*, 107575–107586. [CrossRef]
59. Vilone, G.; Longo, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion* **2021**, *76*, 89–106. [CrossRef]
60. Vilone, G.; Longo, L. Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 615–661. . [CrossRef]