Integrated Spatio-Temporal Deep Clustering (ISTDC) for cognitive workload assessment

Debashis Das Chakladar \(^{a, *}\), Partha Pratim Roy \(^{b, 1}\), Victor Chang \(^{c, 1}\)

\(^{a}\) Department of Electronics and Communication Sciences Unit, Indian Statistical Institute Kolkata, West Bengal 700108, India

\(^{b}\) Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, Uttarakhand 247667, India

\(^{c}\) Department of Operations and Information Management, Aston Business School, Aston University, UK

A B S T R A C T

Traditional high-dimensional electroencephalography (EEG) features (spectral or temporal) may not always attain satisfactory results in cognitive workload estimation. In contrast, deep representation learning (DRL) transforms high-dimensional data into cluster-friendly low-dimensional feature space. Therefore, this paper proposes an Integrated Spatio-Temporal Deep Clustering (ISTDC) model that uses DRL followed by a clustering method to achieve better clustering performance. The proposed model is illustrated using four Algorithms and Variational Bayesian Gaussian Mixture Model (VBGMM) clustering method. Temporal and spatial Variational Auto Encoder (VAE) models (mentioned in Algorithm 2 and Algorithm 3) learn temporal and spatial latent features from sequence-wise EEG signals and scalp topographical maps using the Long short-term memory and Convolutional Neural Network models. The concatenated spatio-temporal latent feature (mentioned in Algorithm 4) is passed to the VBGMM clustering method to efficiently estimate workload levels of n-back task. For the 0-back vs. 2-back task, the proposed model achieves the maximum mean clustering accuracy of 98.0%, and it improves by 11.0% over the state-of-the-art method. The results also indicate that the proposed multimodal approach outperforms temporal and spatial latent feature-based unimodal models in workload assessment.

1. Introduction

Cognitive workload can be defined as a multidimensional construct representing the load that performing a particular task imposes on the learner’s cognitive system \([1]\). The workload level of an operator can be evaluated using subjective or physiological measures. To assess workload levels using traditional subjective measures, self-rating-based methods are used where results mostly depend on an individual subject's honesty. Therefore, an objective measurement based on physiological signals is indispensable. There exist several types of physiological measures that include cardiac activity: Electrocardiography (ECG), respiratory activity, eye activity, and brain activity: Electroencephalography (EEG), Functional magnetic resonance imaging (fMRI), Near-infrared spectroscopy (NIRS), Functional near-infrared spectroscopy (NIRS). Due to the ability to reflect the electrical activities of the cortex, EEG is used as the most effective physiological measure for estimating cognitive workload levels \([2]\). EEG is used in several cognitive applications such as emotion classification \([3,4]\), mental arithmetic tasks \([5]\), n-back tasks \([6]\), and simultaneous multitasking activities \([7]\). Mostly, high-dimensional power spectral density (PSD) features \([8]\), or event-related potential (ERP) features \([9]\) of EEG is used for workload estimation. EEG features have been used in several medical applications. A hybrid framework consisting of a complex brain network and Takagi–Sugeno–Kang fuzzy system has been used to identify Alzheimer’s disease \([10]\). The topological features of functional brain networks have been used for efficient classification. The spectral power of EEG has been used in acupuncture stimulation \([11]\). Apart from PSD, topological graph features such as clustering coefficient, global efficiency, etc have been used in acupuncture manipulation \([12]\). However, the classification model using the unimodal hand-crafted EEG features (temporal: ERP or spectral: PSD) does not give a satisfactory result \([13]\). Multimodal fusion technique can be implemented by combining data of different sensors \([14]\), by fusing different types of features of the same sensor \([15]\) or based on the decision of different unimodal feature extractors \([16]\). Zhang et al. \([17]\) have developed a deep multimodal framework consisting of temporal and spectral EEG features and a deep Convolutional Neural Network (CNN) model to

\(^{*}\) Corresponding author.

E-mail addresses: ddaschakladar@gmail.com (D.D. Chakladar), partha@cs.iitr.ac.in (P.P. Roy), v.chang1@aston.ac.uk, victorchang.research@gmail.com (V. Chang).

\(^{1}\) Senior Member, IEEE.
estimate workload levels from n-back task. The multimodal-CNN model achieved 91.9% classification accuracy. Spatial–temporal autoencoder (STAE) constructed from CNN and Long short-term memory (LSTM) has been used to find the brain dynamics of Alzheimer patients [18]. The STAE model has achieved 96.30% classification accuracy. In [19], authors have merged multimodal features from EEG (PSD and ERP features) and ECG (Heart rate: HR feature) for the n-back experiment. In addition, HR and heart rate variability (HRV) features have been extracted from ECG signals. The fused features from both EEG and ECG sensors were passed to the different classifiers (k-nearest neighbors: kNN, SVM). Among different classifiers, SVM achieves the maximum classification accuracy of 90.6%. Lin et al. [20] have combined EEG (spectral power of EEG bands), and fNIRS (oxygenated hemoglobin: HbO and deoxygenated hemoglobin: HbR) signals to identify workload levels in the lane-deviation driving task. The results showed that an increased concentration of HbO and variation of EEG spectral power for theta, alpha, and beta bands were associated with poor driving performance. However, the feature space becomes large when combining multiple sensor/ modality data and high dimensional feature vectors, often giving poor classification results [21]. Moreover, traditional machine learning or deep learning-based classification methods generally suffer from high computational complexity on large-scale data [22]. To mitigate the computational complexity, the deep representation learning (DRL) method is extensively used alongside clustering to map the input data into feature space in such a way that separation in class becomes more relevant to the problem’s context [23].

Deep neural network architecture such as AutoEncoder (AE) is mainly used in the DRL-based clustering techniques that embed the higher-level features from input data to capture task-specific contextual information. However, due to the complex behavior of EEG signals, AE and Sparse AutoEncoder (SAE)-based deep models [17,24] are unable to identify the underlying structure of the input EEG data. The Variational AutoEncoder (VAE) overcomes this issue by implementing the variational Bayesian method, which improves the scalability of the base network and, consequently, enhances the clustering performance [22]. The VAE model extracts the probabilistic properties from underlying EEG data [25]. Probabilistic properties allow better understanding and inference of underlying cognitive states by identifying the probability distributions of EEG features associated with different cognitive workload levels. In contrast, AE and sparse AE-based deep models are unable to find the underlying probability distributions of EEG features. Moreover, VAE also extracts noise-free localized latent features from input EEG, which improves classification results of EEG tasks [26]. EEG-based applications such as speech recognition [27] or emotion classification [28] have been implemented using the LSTM-based VAE model. VAE model constructed using CNN has been used for motor imagery classification [29]. However, for the large-sized EEG dataset, the computational complexity of deep models is expensive [22]. In contrast, in data clustering, samples (without class labels) are grouped into different clusters by minimizing inter-cluster similarity and maximizing intra-cluster similarity [21]. In the Variational Bayesian Gaussian Mixture Model (VBGMM), an appropriate model selection algorithm has been performed using the posterior distribution of data [30]. VBGMM has been used in EEG-based epileptic seizure detection [31], speaker identification [32] and speech recognition [33]. The VBGMM clustering method leads to better cluster covariance than other clustering methods [21].

Recent DRL-based clustering studies mainly highlighted the image recognition task [34,35], so designing a robust DRL-based clustering method that combines multiple EEG features (temporal and spatial) for workload estimation is of utmost demand. As EEG is suffering from the curse-of-dimensionality issue [21], and traditional classification techniques suffer from large computational complexity for executing large amounts of EEG data, it is essential to propose a DRL-based clustering model that can effectively estimate the workload levels from low-dimensional latent features. The proposed Integrated Spatio-Temporal Deep Clustering (ISTDC) framework is constructed using four algorithms (Algorithm 1–Algorithm 4). The temporal and spatial information from EEG data is captured using LSTM and CNN-based VAE in Algorithm 2 and Algorithm 3, respectively. Outputs of these algorithms (i.e., low-dimensional latent features of temporal and spatial VAEs) are combined into the final latent feature (output of Algorithm 4). The final merged latent feature is passed to the VBGMM clustering method for identifying workload levels. The proposed framework and its relationship with the Algorithms (1–4) are plotted in Fig. 1.

The contributions of this study are mentioned below:
The proposed ISTDC framework is illustrated by four Algorithms (Encode, Temporal VAE, Spatial VAE, and Multimodal integration) followed by a deep clustering method. As the clustering efficiency mostly depends on learned feature representation [21], the VBGM clustering method can effectively classify workload levels using the combined (temporal and spatial) deep latent feature.

For the 0 vs. 2-back task, the proposed model achieves the maximum classification accuracy, and the proposed multimodal model demonstrates superior performance than the unimodal (spatial and temporal) VAE-based clustering approaches by 15.8% and 13.7%, respectively. Furthermore, the multimodal framework exceeds the current deep clustering approaches, improving clustering accuracy by 13.5%. The proposed model is also tested over two other open-access n-back datasets to demonstrate its effectiveness.

(3) Different kinds of comparison studies are performed to evaluate the efficiency of the proposed model. A significant performance improvement of the proposed model is observed for all types of comparisons.

The remainder of the paper is designed as follows. The proposed model is discussed in Section 2. The experimental results of the proposed model are represented in Section 3. A brief discussion is illustrated in Section 4. Finally, in Section 5, the paper is concluded with future work.

2. Methodology

The proposed ISTDC model is divided into three subsections: (A) Dataset and experiment analysis, (B) Integrated spatio-temporal VAE (IST-VAE) model, and (C) Cognitive workload estimation using VBGM. A detailed description of each subsection is mentioned below.

2.1. Dataset and experimental analysis

An open-access public dataset [36] is used for evaluating the proposed model. The dataset contains EEG recordings of 26 subjects (9 males and 17 females, average age of 26.1 ± 3.5 years). EEG data were recorded using 30 EEG electrodes (Fp1, Fp2, AFF5 h, AFF6 h, AFz, F1, F2, FC1, FC2, FC5, FC6, Cz, C3, C4, T7, T8, CP1, CP2, CP5, CP6, Pz, P3, P4, P7, P8, PoZ, O1, O2, TP9 (reference) and TP10 (ground)) at a sampling rate of 1,000 Hz. Then, the raw EEG was filtered with a passband of 1–40 Hz to remove high-frequency noise from EEG. Next, the Independent component analysis (ICA) was applied to remove artifacts (ocular, cardiac) from EEG.

The experimental n-back dataset includes three sessions, where each session is divided into three series. In the experiment, nine series of n-back tasks are performed for each participant. Each session consists of 20 trials; thus the experiment includes a total of 180 (20 trials x 3 series x 3 sessions) trials for each n-back task. A single series was composed of a 2 s instruction showing the type of the task (0-, 2- or 3-back), a 40 s task period, and a 20 s rest period. In the n-back task, participants need to identify the letter/digit presented n trials earlier in the sequence. In the task state, a digit (i.e., stimulus) randomly appeared on a screen for 2 s. In the rest state, subjects were asked to keep their eyes closed without performing any task.

2.2. Integrated spatio-temporal VAE model

The DRL-based method, A non-linear mapping function $f_x : X → Z$ converts the high-dimensional input $X \in \mathbb{R}^d$ into low-dimensional embedded feature space $(Z \in \mathbb{R}^k$, where $k \ll d)$. The input and reduced dimensions are represented as $d$ and $k$, respectively. Due to better feature learning capabilities and function approximation properties than other deep neural networks (DNN) methods, AEs are widely used for parameterization of $f_x$. In VAE, the regularization effect on the latent variables overcomes the overfitting issue of AE. The entire process of the IST-VAE model is illustrated using four algorithms. Algorithm 1 depicts the encoding process of input data into a latent variable. Algorithm 2 and Algorithm 3 illustrate the construction process of the encoder in both VAEs. The two encoders ($TEncoder$ in Algorithm 2 and $SEncoder$ in Algorithm 3) of VAEs have extracted the corresponding latent features from inputs. Two encoders are built using LSTM and CNN models. Algorithm 4 concatenates the latent features (SpLat and TemLat) obtained from $TEncoder$ and $SEncoder$ and passed to the VBGM clustering method for workload classification. In the experiment, the sequence-specific (two seconds) EEG signals and topographical images were passed as input to the temporal and spatial VAEs. VAE extracts the localized features from topographical images of EEG [37].

Morlet wavelet features are useful to identify the localized changes of signal for different frequency components over time [38]. It can be noted that Common Spatial Pattern (CSP), spectral power and mean absolute band power features of the alpha band of EEG can effectively classify workload levels from different cognitive tasks such as ‘simultaneous capacity-based multitasking activity [7]’, ‘mental arithmetic [39]’ and ‘modified Sternberg task [40]’ respectively. So, for different cognitive tasks, it can be observed that the alpha band of EEG can effectively classify the workload states than other EEG bands. Hence, the Morlet wavelet features are estimated based on the best EEG band (i.e., alpha band) for workload estimation. In the 40 s of the specific task (0/2/3-back) period in the experiment, each digit displays for 2 s; thus, 20 scalp topographical images ($\frac{20 \times 9}{2} = 90$) are extracted for each type of task. So, 20 task-specific topographical images are constructed from one series. Thus, for nine such series, a total of 180 ($20 \times 9$) images are constructed for each subject. Scalp topographical maps are generated from the sequence (time interval of 2 s) specific Morlet wavelet features. The topographical map refers to the spatial distribution of brain electrical activity for a specific frequency band (here, alpha band). The proposed spatial VAE model takes the topographical images (size: $80 \times 60 \times 3$, where $80$, $60$, and 3 are represented as height, width, and number of image channels) as input (line 2 of Algorithm 3). The construction of the CNN-based VAE model (Spatial VAE) is illustrated in line 2–11 of Algorithm 3. In the spatial VAE model, 16 filters of size: $3 \times 3$ are applied on inputs, and outputs of the first convolutional layer are passed to the first Max pooling layer (width: 4 and stride: $2 \times 2$). The output of the first Max-pooling layer is passed to the second convolutional layer of 24 filters (size: $5 \times 5$). The output of the second convolutional layer is passed to the second Max pooling layer (width: 2 and stride: $1 \times 1$). The second Max pooling layer is followed by a Flatten layer. Finally, a dense layer of eight neurons (best latent feature dimension, refer to Fig. 4(a)) completes the model configuration. The construction process of temporal VAE using stacked LSTM layers is depicted in line 4 to 17 of Algorithm 2. Two LSTM layers (6, 12, and 18 units, respectively) are used to build the proposed temporal VAE, where the output of the first LSTM layer is passed as an input to the second LSTM layer. Next, the output of the second LSTM layer is fed into the third LSTM layer. A batch normalization and dropout layer (dropout rate: 0.2) are appended after each LSTM layer. A dense layer of eight neurons (best latent feature dimension, refer to Fig. 4(a)) followed by the third dropout layer (i.e., d3) completes the model configuration. For both VAEs (Algorithm 2 and Algorithm 3), the input is encoded into latent variables (TemLat and SpLat) through the Encoding algorithm (Algorithm 1). The output dimension of both VAEs is maintained at the same value (i.e., eight) to combine the resultant latent features of both temporal and spatial VAEs into a shared embedded space. Fig. 1 shows the configuration of temporal and spatial VAE. The merging/concatenation process of two latent features is performed in line 1 in Algorithm 4.
2.3. Cognitive workload estimation using VBGMN

The concatenated latent feature vector (output of Algorithm 4) is passed to the VBGMN clustering for workload classification. In the Variational BGMN approach, approximate posterior distributions can be effectively determined using the variational inference algorithm while retaining the advantages of the Bayesian approach [41]. Since the probabilistic variables of each category (c) are statistically independent, the approximate posterior distribution of the VBGMN model can be expressed as:

\[ p(\theta, Z | X, m) = \prod p(\theta_c | X_c, m)p(Z_c | X_c, m) \]  \hspace{1cm} (1)

where, \( Z \) is the hidden variable of model \( m \). \( p(Z | X, m) \) is the approximate posterior distribution of model parameters, and \( p(\theta) | X, m \) refers to the solutions of VB posterior distribution. The VBGMN clustering method is related to two parameters: (a) “prior type” (Dirichlet process or Dirichlet distribution) and (b) “weight_concentration_prior”, which refers to the distribution of weights to each of the components based on prior type [30]. The output components or clusters can be inferred using this “weight” parameter. Here, the weight parameter value is selected as \( 1e+2 \) and Dirichlet distribution is chosen as the prior type of the model. This extra parameterization is necessary for variational inference, but for the prior type “Dirichlet process”, the inference process may be slower.

Here, the performance evaluation of VBGMN clustering is performed using three metrics, namely unsupervised clustering accuracy (Acc), Normalized Mutual Information (NMI), and Rand Index (RI). Acc (2) refers to the best matching between cluster assignments from the clustering method (\( t_i \)) and ground truth labels (\( y_i \)).

\[ Acc = \max_{\hat{t}} \frac{1}{n} \sum_{i=1}^{n} I(y_i = m(\hat{t})) \] \hspace{1cm} (2)

where \( n \) is the number of samples and \( m \) ranges overall possible one-to-one mappings between clusters and labels. NMI (3) refers to the reduction of entropy information for cluster assignments with respect to the known ground truth labels.

\[ NMI(y,m) = \frac{I(y,m)}{\frac{1}{2}[H(y) + H(m)]} \] \hspace{1cm} (3)

where the ground truth labels and cluster assignment are denoted by \( y \) and \( m \), respectively. Mutual information between \( y \) and \( m \) is represented by \( I \), and entropy is denoted by \( H(.) \). The Rand Index (RI) checks the similarity between original and predicted class labels, represented as follows:

\[ RI = \frac{TP + TN}{TP + FP + TN + FN} \] \hspace{1cm} (4)

where, TP, TN, FP, and FN signify true positive, true negative, false positive and false-negative rates, respectively.

3. Results

This section is divided into five subsections: (A) behavioral results, (B) integrated spatio-temporal VAE model analysis, (C) clustering results & performance analysis, (D) computational complexity analysis, and (E) comparative analysis. The detailed analysis of each section is mentioned below.

3.1. Behavioral results

As the task complexity level is related to the value of \( n \) (\( n = 2, 3 \)) in the \( n \)-back task, the subject needs to remember more numbers of previous letters with the higher value of \( n \). The Friedman test has been performed to check whether the means of each workload level are equal or not. The result (\( p < 0.05, \chi^2 = 32.077, df = 30 \)) indicates that a significant difference exists in mean values of different workload levels. After the Friedman test, the Dunn–Bonferroni Posthoc test was conducted to determine the difference between each pair of workload levels. The Posthoc test (Table 1) validates that the mean differences between each pair of workload levels are statistically significant (\( p \) value < 0.05).
training, the Adam optimizer (learning rate of 0.02) is used. A batch
proposed IST-VAE. Each deep model is trained on 100 epochs. During
VAE with three-layered LSTM (Model 1 in Fig. 2) is selected for the
than the performance of three-layered LSTM model. Thus, the temporal
mentioned single and two-layered LSTM models, which are much lesser
clustering accuracy of 0.526 and 0.583, respectively, using these above-
workload condition (i.e., cond 4), the IST-VAE model achieves the mean
not changed while computing the combined latent features. For high
the configuration of the best spatial VAE (in Model 1 of Fig. 2) is
among the different combinations of VAE and temporal VAE models are shown in Fig. 2. Different configurations of temporal and spatial VAE models are constructed using Algorithm 2 and Algorithm 3 by changing the layer-wise LSTM units or layer-wise filters of the CNN model. The training losses (KL divergence) of different model combinations (temporal-spatial VAE) are shown in Fig. 2. Since the first temporal and spatial VAE model combination (Model 1: Fig. 2(a)) incurs the lowest KL divergence loss, this configuration is selected for the proposed IST-VAE model. The temporal VAE of the proposed model is already analyzed with different combinations of LSTM layers (Model 1: 3 layers, Model 2: 2 layers, and Model 3: 4 layers) in Fig. 2. However, for generalization purposes, the temporal VAE is also experimented with one-layer LSTM model (L128) and another two-layered (L12-L18) LSTM model. Here, \( L \) represents the number of LSTM units. For the two differently configured temporal VAEs, the configuration of the best spatial VAE (in Model 1 of Fig. 2) is not changed while computing the combined latent features. For high workload condition (i.e., cond 4), the IST-VAE model achieves the mean clustering accuracy of 0.526 and 0.583, respectively, using these above-mentioned single and two-layered LSTM models, which are much lesser than the performance of three-layered LSTM model. Thus, the temporal VAE with three-layered LSTM (Model 1 in Fig. 2) is selected for the proposed IST-VAE. Each deep model is trained on 100 epochs. During training, the Adam optimizer (learning rate of 0.02) is used. A batch size of 32 was chosen during the training of the IST-VAE model. The hyperparameters (learning rate, batch size, number of epochs, number of hidden layers, number of neurons/layers etc.) of the IST-VAE model are tuned using the Random search method [42]. Like Grid search, the Random search does not exhaustively examine all conceivable parameter combinations; instead, it selectively draws hyperparameter values from predefined distributions in a random manner. Thus, this approach discovers effective hyperparameters quicker than the exhaustive grid search method.

Gradient-Class Activation Mapping (Grad-CAM) visualization technique utilizes gradients from the final convolutional layer to generate a coarse localization map, indicating the spatial areas where electrodes are positioned [43]. The feature map obtained by Grad-CAM from the last convolution layer contains high-level features and spatial information [43]. Thus, spatial localization maps extracted from the last convolution layer of the proposed spatial VAE model highlight the spatial information of scalp topographical images. Three randomly selected topographical images of each workload level and their spatial localization feature maps (obtained from Grad-CAM) are shown in Fig. 3. The brighter color in the localized map indicates the activated/informative region in the brain for the workload class. For the 0-back class, high activation is observed in the right frontal and left parietal brain lobes, whereas a high activation is observed only in the left parietal brain lobe for the 2-back task. Only the right frontal brain area is activated for the complex task (3-back).

<table>
<thead>
<tr>
<th>Sample 1-Sample 2</th>
<th>Mean difference</th>
<th>Std. error</th>
<th>( p ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-back-0-back</td>
<td>1.056</td>
<td>0.297</td>
<td>0.007</td>
</tr>
<tr>
<td>3-back-2-back</td>
<td>1.114</td>
<td>0.321</td>
<td>0.011</td>
</tr>
<tr>
<td>3-back-0-back</td>
<td>1.127</td>
<td>0.336</td>
<td>0.012</td>
</tr>
</tbody>
</table>

3.2. Integrated spatio-temporal VAE model analysis

This section is further divided into three subsections: (1) selection of the VAE model & Spatial feature analysis, (2) effects of the latent dimension of VAEs, and (3) effects of optimizer and learning rate. The detailed discussions of each subsection are mentioned below.

3.2.1. Selection of the VAE model & spatial feature analysis

The construction of deep VAE models is illustrated in Algorithm 2 and Algorithm 3. Both algorithms are used only encoder parts of VAEs. Two types of losses are associated during the training process of VAE: reconstruction loss and Kullback–Leibler (KL) divergence. However, as the proposed model only uses the encoder’s output, only KL divergence loss is considered. The KL divergence tends to regularize the organization of the latent space by making the distributions returned by the encoder closer to a standard normal distribution. Hence, the objective is to select the optimal VAE model such that the KL divergence between the distribution of the encoder and input will be minimized. KL divergence values of different temporal and spatial VAE models are shown in Fig. 2. Different configurations of temporal and spatial VAE models are constructed using Algorithm 2 and Algorithm 3 by changing the layer-wise LSTM units or layer-wise filters of the CNN model. The training losses (KL divergence) of different model combinations (temporal-spatial VAE) are shown in Fig. 2. Since the first temporal and spatial VAE model combination (Model 1: Fig. 2(a)) incurs the lowest KL divergence loss, this configuration is selected for the proposed IST-VAE model. The temporal VAE of the proposed model is already analyzed with different combinations of LSTM layers (Model 1: 3 layers, Model 2: 2 layers, and Model 3: 4 layers) in Fig. 2. However, for generalization purposes, the temporal VAE is also experimented with one-layer LSTM model (L128) and another two-layered (L12-L18) LSTM model. Here, \( L \) represents the number of LSTM units. For the two differently configured temporal VAEs, the configuration of the best spatial VAE (in Model 1 of Fig. 2) is not changed while computing the combined latent features. For high workload condition (i.e., cond 4), the IST-VAE model achieves the mean clustering accuracy of 0.526 and 0.583, respectively, using these above-mentioned single and two-layered LSTM models, which are much lesser than the performance of three-layered LSTM model. Thus, the temporal VAE with three-layered LSTM (Model 1 in Fig. 2) is selected for the proposed IST-VAE. Each deep model is trained on 100 epochs. During training, the Adam optimizer (learning rate of 0.02) is used. A batch size of 32 was chosen during the training of the IST-VAE model. The hyperparameters (learning rate, batch size, number of epochs, number of hidden layers, number of neurons/layers etc.) of the IST-VAE model are tuned using the Random search method [42]. Like Grid search, the Random search does not exhaustively examine all conceivable parameter combinations; instead, it selectively draws hyperparameter values from predefined distributions in a random manner. Thus, this approach discovers effective hyperparameters quicker than the exhaustive grid search method.

Gradient-Class Activation Mapping (Grad-CAM) visualization technique utilizes gradients from the final convolutional layer to generate a coarse localization map, indicating the spatial areas where electrodes are positioned [43]. The feature map obtained by Grad-CAM from the last convolution layer contains high-level features and spatial information [43]. Thus, spatial localization maps extracted from the last convolution layer of the proposed spatial VAE model highlight the spatial information of scalp topographical images. Three randomly selected topographical images of each workload level and their spatial localization feature maps (obtained from Grad-CAM) are shown in Fig. 3. The brighter color in the localized map indicates the activated/informative region in the brain for the workload class. For the 0-back class, high activation is observed in the right frontal and left parietal brain lobes, whereas a high activation is observed only in the left parietal brain lobe for the 2-back task. Only the right frontal brain area is activated for the complex task (3-back).
6

3.2. Effects of latent dimension of VAEs

The image and EEG signal dimensions of the proposed IST-VAE model are $M_1: 80 \times 60 \times 3 = 14400$ and $M_2 = 30$ (no. of EEG channels), respectively. As the outputs of two latent features (from temporal and spatial VAEs) are merged into an embedded feature space, thus, the resultant latent dimension ($M$) in the embedded feature space should be lower than $M_2$ (as $M_2 < M_1$). For simplicity, the experiment is analyzed between 0-back vs 2-back workload levels. The proposed model is trained for all subjects using the leave-subject-out (LSO) test across different latent dimensions. In Fig. 4(a), the average clustering performance is evaluated based on the multimodal latent feature (output of Algorithm 4) of the IST-VAE model with different latent dimensions (by changing the different $\epsilon$ps value of Algorithm 1). It can be noted that the maximum clustering performance is achieved with $M = 8$. Clustering performance varies with different latent dimensions of Temporal and Spatial VAEs (variable $M_2$ in Algorithm 2 and $M_1$ in Algorithm 3). The larger value of $M$ may preserve more information but simultaneously minimize the discrimination power of learned representation, which decreases the clustering performance [44]. The same latent dimension (i.e., $M = 8$) is maintained for other workload conditions.

3.2.3. Effects of optimizer and learning rate

In this section, hyperparameters of the proposed IST-VAE model are tuned to obtain the best clustering performance. The convergence of deep neural networks largely depends on selecting the proper optimizer and learning rate ($lr$). The VAE model is trained using the leave-subject-out (LSO) method to avoid the overfitting issue. The proposed VAE model is trained with different optimizers and its learning rate ($lr$). Next, the clustering accuracy is computed for each subject of the test set. The process is repeated for all the subjects. The mean clustering accuracy of different optimizers (adam, stochastic gradient descent: SGD, adadelta, and RMSprop) and their learning rates ($lr$) (0.01 to 0.1) is reported in Fig. 4(b). It can be noted that the Adam optimizer with the $lr$ of 0.02 achieves the maximum clustering accuracy. So, the Adam optimizer with the same $lr$ is used for other workload conditions.

However, the experiment result of the 0-back vs. 2-back workload condition is presented for simplicity.

3.3. Clustering results & performance analysis

This section is divided into two subsections: (1) Output analysis of the VBGMM clustering method and (2) performance analysis of VBGMM based on other datasets. A detailed description of each section is mentioned below.

3.3.1. Output analysis of the VBGMM clustering method

This section highlights the output analysis of the VBGMM clustering method and subject-wise clustering results of different workload levels. The merged latent feature (i.e., the combination of temporal and spatial latent features) created in Algorithm 4 is used as input for clustering. The LSO experiment (the training and testing set consists of entirely different subjects) is performed to evaluate the subject-wise variation across various workload levels. In the LSO experiment, the EEG signals and images of scalp topographical maps of 25 subjects are used as a training set. The testing set comprised the remaining subject’s topographical maps and EEG data. After the LSO experiment, the results are reported in Fig. 4(a), which shows average clustering performance compared to different latent dimensions. The best performance is achieved with the latent dimension of 8. However, the clustering results and performance analysis are presented below.

### Table 2

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Cond 1</th>
<th>Cond 2</th>
<th>Cond 3</th>
<th>Cond 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub 1</td>
<td>0.988</td>
<td>0.869</td>
<td>0.882</td>
<td>0.771</td>
</tr>
<tr>
<td>Sub 2</td>
<td>0.965</td>
<td>0.802</td>
<td>0.723</td>
<td>0.602</td>
</tr>
<tr>
<td>Sub 3</td>
<td>0.989</td>
<td>0.823</td>
<td>0.798</td>
<td>0.781</td>
</tr>
<tr>
<td>Sub 4</td>
<td>0.981</td>
<td>0.844</td>
<td>0.811</td>
<td>0.689</td>
</tr>
<tr>
<td>Sub 5</td>
<td>0.992</td>
<td>0.881</td>
<td>0.827</td>
<td>0.708</td>
</tr>
<tr>
<td>Sub 6</td>
<td>0.979</td>
<td>0.913</td>
<td>0.933</td>
<td>0.856</td>
</tr>
<tr>
<td>Sub 7</td>
<td>0.987</td>
<td>0.893</td>
<td>0.922</td>
<td>0.898</td>
</tr>
<tr>
<td>Sub 8</td>
<td>0.952</td>
<td>0.833</td>
<td>0.721</td>
<td>0.779</td>
</tr>
<tr>
<td>Sub 9</td>
<td>0.985</td>
<td>0.923</td>
<td>0.907</td>
<td>0.658</td>
</tr>
<tr>
<td>Sub 10</td>
<td>0.978</td>
<td>0.883</td>
<td>0.743</td>
<td>0.593</td>
</tr>
<tr>
<td>Sub 11</td>
<td>0.981</td>
<td>0.923</td>
<td>0.899</td>
<td>0.738</td>
</tr>
<tr>
<td>Sub 12</td>
<td>0.992</td>
<td>0.995</td>
<td>0.949</td>
<td>0.689</td>
</tr>
<tr>
<td>Sub 13</td>
<td>0.977</td>
<td>0.889</td>
<td>0.836</td>
<td>0.934</td>
</tr>
<tr>
<td>Sub 14</td>
<td>0.974</td>
<td>0.936</td>
<td>0.811</td>
<td>0.941</td>
</tr>
<tr>
<td>Sub 15</td>
<td>0.978</td>
<td>0.973</td>
<td>0.944</td>
<td>0.863</td>
</tr>
<tr>
<td>Sub 16</td>
<td>0.985</td>
<td>0.962</td>
<td>0.884</td>
<td>0.778</td>
</tr>
<tr>
<td>Sub 17</td>
<td>0.984</td>
<td>0.923</td>
<td>0.889</td>
<td>0.924</td>
</tr>
<tr>
<td>Sub 18</td>
<td>0.972</td>
<td>0.856</td>
<td>0.923</td>
<td>0.899</td>
</tr>
<tr>
<td>Sub 19</td>
<td>0.986</td>
<td>0.902</td>
<td>0.823</td>
<td>0.952</td>
</tr>
<tr>
<td>Sub 20</td>
<td>0.985</td>
<td>0.853</td>
<td>0.771</td>
<td>0.851</td>
</tr>
<tr>
<td>Sub 21</td>
<td>0.989</td>
<td>0.829</td>
<td>0.745</td>
<td>0.798</td>
</tr>
<tr>
<td>Sub 22</td>
<td>0.991</td>
<td>0.923</td>
<td>0.887</td>
<td>0.923</td>
</tr>
<tr>
<td>Sub 23</td>
<td>0.986</td>
<td>0.905</td>
<td>0.899</td>
<td>0.911</td>
</tr>
<tr>
<td>Sub 24</td>
<td>0.977</td>
<td>0.825</td>
<td>0.786</td>
<td>0.836</td>
</tr>
<tr>
<td>Sub 25</td>
<td>0.983</td>
<td>0.893</td>
<td>0.843</td>
<td>0.893</td>
</tr>
<tr>
<td>Sub 26</td>
<td>0.979</td>
<td>0.901</td>
<td>0.916</td>
<td>0.928</td>
</tr>
</tbody>
</table>

Mean: 0.980 0.890 0.848 0.821 0.750 0.738 0.899 0.833 0.805 0.789 0.797 0.796

Fig. 4. Evaluation of clustering performance: (a) based on different latent dimensions, (b) based on different optimizers and their learning rates for workload condition 0-back vs. 2-back. Note: Mean clustering accuracy (MACC), Mean Normalized Mutual Information (MMNI), and Mean Rand Index (MRI).
the trained latent feature vector of VAE is used for the training of VBGMM, and the predicted latent feature vector is used to evaluate clustering performance. The training parameters of VBGMM are as follows: maximum iterations = 150, weight_concentration_prior = 0.01, weight_concentration_prior_type = Dirichlet process, covariance_type = full, and mean_precision_prior = 0.1. The LSO experiment is performed for all subjects with different train-test sets, and the result is shown in Table 2. The average clustering performance result (MACC and MNMI) for all workload conditions is displayed in Table 3.

Table 3

<table>
<thead>
<tr>
<th>Conditions</th>
<th>IST-VAE MACC</th>
<th>Spatial VAE MACC</th>
<th>Temporal VAE MACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConD 1</td>
<td>0.980</td>
<td>0.890</td>
<td>0.822</td>
</tr>
<tr>
<td>ConD 2</td>
<td>0.821</td>
<td>0.750</td>
<td>0.689</td>
</tr>
<tr>
<td>ConD 3</td>
<td>0.899</td>
<td>0.833</td>
<td>0.726</td>
</tr>
<tr>
<td>ConD 4</td>
<td>0.789</td>
<td>0.737</td>
<td>0.623</td>
</tr>
</tbody>
</table>

In manifold learning, the high-dimensional data can be represented in low-dimensional space for better visualization. Through the non-linear mapping operation, uniform manifold approximation and projection (UMAP) transforms higher-dimensional data into a lower-dimensional manifold with the fastest execution time [45]. The UMAP cluster plots (Fig. 5) of the VBGMM clustering method are shown for three different latent dimensions (6, 8, and 10) of the IST-VAE model. The different-sized latent dimensions are obtained by changing the $\epsilon$ variable in Algorithm 1. The clustering quality is highly dependent on the learned representation of data [21]. It can be noted that the cluster representation for latent dimension 8 is better than other latent dimensions (6, 8, and 10) of the IST-VAE model. The different-sized latent dimensions are obtained by changing the $\epsilon$ variable in Algorithm 1. The clustering quality is highly dependent on the learned representation of data [21]. It can be noted that the cluster representation for latent dimension 8 is better than other latent dimensions (6, 8, and 10) of the IST-VAE model.

3.3.2. Performance analysis of VBGMM based on other datasets

For generalization, the proposed model has been implemented to two publicly open-access $n$-back datasets [46,47]. Each open-access dataset has a different workload condition. The LSO experiment is performed for each dataset, and the VBGMM clustering result (Table 4) is produced based on the MACC and MNMI matrices. It can be observed that the proposed model achieves the best clustering performance (MACC: 0.923, MNMI: 0.878) for the workload level (0-back vs. 1-back) on the dataset [46]. Therefore, it can be stated that the proposed model is capable of accurately estimating different workload conditions across various datasets.

Table 4

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Dataset 1 [46] MACC</th>
<th>MNMI</th>
<th>Dataset 2 [47] MACC</th>
<th>MNMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1 back</td>
<td>0.923</td>
<td>0.878</td>
<td>0-1 back</td>
<td>0.911</td>
</tr>
<tr>
<td>0-2 back</td>
<td>0.917</td>
<td>0.852</td>
<td>0-2 back</td>
<td>0.802</td>
</tr>
<tr>
<td>1-2 back</td>
<td>0.825</td>
<td>0.773</td>
<td>1-2 back</td>
<td>0.769</td>
</tr>
<tr>
<td>0-1-2 back</td>
<td>0.703</td>
<td>0.683</td>
<td>0-1-2 back</td>
<td>0.733</td>
</tr>
</tbody>
</table>
### 3.4. Computational complexity analysis

This section discusses the computational complexity of the proposed model. In the proposed model, the LSTM and CNN-based VAE models are used to extract temporal and spatial latent features from inputs. The time complexity of the single LSTM network is \( O(N_s D + M) \), where \( N_s \) and \( D \) are the time step and weights, respectively [48]. Here, each weight is associated with a single node. The computational complexity of the CNN model is \( O(N W) \), where \( N \times N \) is the input image size [49]. The concatenation of latent features (Algorithm 4) and the encoding operation (Algorithm 1) takes \( O(1) \) time. The complexity of the VBGMM clustering algorithm is \( O(s D^3) \), where \( s \) and \( D \) are the numbers of iterations and feature dimensions [50]. Therefore, the overall complexity of the proposed model is \( O(W^2 + N + s D^3) \).

### 3.5. Comparison study

Several comparison studies are performed in this section, namely, (1) a comparison among deep models for cluster analysis, (2) a comparison among clustering methods, (3) a comparison among deep clustering methods, and (4) a comparative analysis among existing studies. The detailed discussions of each subsection are mentioned below.

#### 3.5.1. Comparison among deep models for cluster analysis

Table 5 compares the VBGMM clustering results using the proposed multimodal latent feature and other latent feature estimation methods, such as PCA and deep latent features (extracted from different AE-based deep models). In Convolutional autoencoder (CAE) and LSTM autoencoder (LAE), clustering performance is evaluated using deep spatial and temporal latent features (i.e., outputs of the CNN and LSTM-based encoders), respectively. The best LSTM and CNN model configuration (refer to Fig. 2(a)) is used in LAE and CAE models. In the denoising autoencoder (DAE), noise is added to the input topographical image with the optimal noise factor of 0.8 [44]. Then, the reconstructed robust (noise-free) spatial feature (from the decoder of DAE) is used as a latent feature for clustering. For Sparse Autoencoder (SAE), an L2 regularizer was utilized with a weight decay factor of 0.003 [51]. The autoencoder model only consists of dense layers. The encoder part is only utilized for extracting latent features in all the autoencoder models except for the DAE. The configuration of all autoencoders is as follows: SAE/CAE → D(128)-D(32)-D(8), DAE → C16(3,3)-M(2)-C32(3,3)-M(2)-DC32(3,3)-U(2)-DC16(3,3)-U(2)-DC8(3,3), CAE → C16(3,3)-M(4)-C24(5,5)-M(2)-F-D(8), LAE → L(6)-L(12)-L(18)-D(8), where \( C(k, k) \): convolutional layer with filters \( (k \times k) \), \( DC \): deconvolution, \( M(n) \): max-pooling (width), \( F \): flatten, \( D \): dense, \( U \): upsampling, \( L \): LSTM layer. The CAE, DAE, SAE, and AE used the topographical images, whereas LAE used the preprocessed EEG signal for training.

It can be identified that the proposed model achieves the best clustering results among all the AE-based models across all the workload conditions.

#### 3.5.2. Comparison among clustering methods

This section highlights the comparison between the proposed multimodal deep clustering method and some popular deep clustering methods for the experimental dataset. The spatial topographical images are only used in the image-based deep clustering methods. The comparison is shown in Table 7. As the proposed model captures both the temporal and spatial features for clustering, it enhances the cluster quality. For all of the experimental conditions, the proposed deep clustering performs better than other deep clustering methods.

#### 3.5.3. Comparison analysis among existing studies

In this section, the comparison is performed between the proposed model and the studies that used the same experimental dataset [36]. The comparison (refer to Table 8) is performed based on the methods and results of those studies. It can be concluded that the low-dimensional latent feature performs better than the traditional EEG features (PSD, ERD) for the same experimental dataset. Thus, the proposed model achieves better accuracy than other models.

### 4. Discussion

This study combines a multimodal deep VAE model and the VBGMM clustering method for workload classification. In this paper, the VBGMM-based probabilistic clustering method is employed since it emerges as a suitable choice for addressing the stochastic nature of human responses [59]. The performance of a proposed IST-VAE model largely depends on the VAE model structure and hyperparameters. The proposed multimodal deep clustering is evaluated with the three different workload conditions with varying degrees of task complexity. The proposed model achieves 98.0% and 99.2% as the best mean
clustering accuracy and subject-wise clustering accuracy, respectively (refer to Table 2). The performance of the proposed deep IST-VAE framework also outperforms unimodal-based VAE models (Table 3) and other AE-based deep models (Table 5). The result indicates that the fusion of low dimensional temporal and spatial EEG features can effectively identify the workload levels of the n-back task. For generalization purposes, the proposed model is examined on two open-access n-back datasets (refer to Table 4), and promising results are found for both datasets. According to comparison studies, the proposed VBGMM clustering approach performs better than traditional well-known clustering methods (refer to Table 6). The proposed multimodal-based clustering method can efficiently classify the different workload levels. Thus, the proposed model can be effectively used to identify human behavior while accomplishing EEG-based cognitive tasks.

5. Conclusion & future work

In this paper, two modalities (temporal and spatial) of EEG are fused to estimate the workload levels of a participant efficiently. Four distinctive algorithms (Encode, Temporal VAE, Spatial VAE, and Multimodal integration) are developed and used to build the multimodal workload estimation framework. The latent feature creation process for the temporal and spatial encoders is performed by Algorithm 1. The selection of a proper latent feature dimension is of utmost importance in DRL-based clustering applications, as changing the latent feature’s dimension can affect the clustering performance. The clustering performance with different latent dimensions is plotted in Fig. 4(a). In Algorithm 2 and Algorithm 3, both temporal and spatial encoders (using deep VAE) extract the layer-wise temporal/spatial meaningful information for classification. Moreover, the LSTM and CNN-based VAEs used in Algorithm 2 and Algorithm 3 are applicable for sequence learning contexts with long-term dependency problems [17]. As the extracted low-dimensional latent feature contains significant information related to the problem context, merging both of these feature vectors into a common embedded feature vector highlights more relevant information than the unimodal approach. As the clustering efficiency especially depends on learned feature representation [21], thus, the VBGMM clustering method can efficiently estimate workload-specific clusters using the merged latent feature (output of Algorithm 4) having the temporal and spatial knowledge of input data.

The experimental dataset consists of the EEG and NIRS data of 26 subjects; however, the experiment is conducted only on EEG signals. In the near future, sensor-level multimodal capabilities (EEG and NIRS) can be implemented in the proposed model to enhance its performance.

Declaration of competing interest

This is hereby certify that the paper is original, neither the paper nor a part of it is under consideration for publication anywhere else. Also, we have no conflicts of interest to disclose.

Data availability

The authors do not have permission to share data.

Acknowledgment

Part of Prof Chang’s work is partly supported by VC Research, UK (VCR0000221). We confirm that the manuscript has been read and approved by all named authors.

References


CRediT authorship contribution statement

Debashis Das Chakladar: Conceptualization, Methodology, Data curation, Software, Writing – original draft. Partha Pratim Roy: Writing – review & editing, Supervision. Victor Chang: Resources, Writing – review & editing, Supervision.
D.D. Chakladar et al.

Biomedical Signal Processing and Control 89 (2024) 105703