


Recognizing Cognitive Load by a Hybrid Spatio-Temporal Causal Model from Multivariate Physiological Data

Zirui Yong¹, Guoxin Su², Xiaohu Li¹, Lingyun Sun³, Zejian Li³, and Li Liu(¹)

¹ School of Big Data & Software Engineering, Chongqing University, China
{yongzirui,xhlee,dcsliuli}@cqu.edu.cn

² School of Computing and Information Technology, University of Wollongong, Australia guoxin@uow.edu.au

³ International Design Institute, Zhejiang University, China
{sunly,zejianlee}@zju.edu.cn

Abstract. Cognitive load recognition is challenging due to the inherent diversity and causality of multivariate physiological changes, with each of its instances having its own style of physiological events and their spatio-temporal causal dependencies. This leads us to define a hybrid model that employs Granger causality (GC) and Gramian angular difference fields (GADF) to discover diverse varieties of multivariate physiological events. In particular, our model introduces a GC network to explicitly characterize the unique temporal causal configurations of a particular cognitive state as a variable number of nodes and links. In addition, GADF maps are constructed to capture the inherit spatio-temporal dependency among multivariate signals in a 2D structural space. A capsule network is designed to merge these two heterogenous types of features together in a uniform way, and as a result, all local causal and spatio-temporal dependencies are globally consistent. Empirical evaluations on one benchmark dataset and two in-house datasets collected by ourselves in virtual reality learning environment suggest our model significantly outperforms the state-of-the-art approaches.

Keywords: Cognitive load recognition · Physiological signal · Granger causality · Gramian angular difference fields

1 Introduction

Cognitive load recognition, aiming to estimate the amount of an individual’s mental labor when a specific task is imposed on her/his cognitive system [21], has become an active field, given its role in facilitating a broad range of applications. Although psychological experiment-based approaches are becoming mature to estimate cognitive load by adopting various subjective scales, they are still limited to obtain the objective states of cognitive load changes in real time. Since an individual’s cognitive load state is often accompanied by the changes

of physiological characteristics such as EEG, ECG, EMG, blood pressure and respiration, it is possible to achieve a deeper understanding of the correlations between long-term measurements of physiological features and cognitive loads. The main focus of this work is on causal learning of multiple physiological features, since a fundamental assumption for research on cognitive load assessment is the causal relationship between the physiological characteristics and cognitive load states.

Despite being a very challenging problem, in recent years there has been a rapid growth of interest in physiologically-based cognitive load recognition. One popular paradigm might be that of the knowledge-driven approaches, which are capable of representing rich relations among physiological events. These approaches are often semantically clear, logically elegant, and easy to interpret. However, physiological features and their causal relations need to be manually defined and extracted, and subsequently they are limited to scale up. For instance, an alarm of a physiological event (e.g., heart rate deviation from a normal range) is triggered by setting a threshold obtained from the psychological domain knowledge or expert experience. It could be rather difficult to handcraft all the signs of features accurately for many practical scenarios where such knowledge embedded in signals are intricate. In addition, these knowledge-driven models are sensitive to sensor noise or body movement, which occurs frequently when performing a task.

On the other hand, data-driven approaches, especially the deep learning-based models, which may overcome the aforementioned shortcomings by automating feature extraction from raw physiological signals. As the fundamental issue in machine learning, current techniques are becoming mature to analyze physiological time series. We refer interested readers to a recent comprehensive review of varied representative physiological data-driven algorithms [20]. With the great success being achieved, these data-driven models are capable of handling an astonishing number of correlations between features and are often robust to errors caused by incorrect physiological detection. However, these conventional approaches have the assumption that physiological features are independent without taking into account the causality between them. Their results are hard to interpret, and therefore, they are rather limited in further uncovering rich cause-effect relationships among features. For instance, the variation of the amplitudes of ECG P-QRS-T waves (P-QRS-T) or the EEG based zero-phase phase-locking value (PLV) is the reaction of *high load state* in cognitive processes [5]. In fact, most of existing data-driven models may find that there is a heavy correlation between P-QRS-T and PLV but unfortunately cannot discover the further interpretation that the *high load state* is the common cause of these two symptoms, which leads to their extrinsic association. As a result, it could be rather difficult to examine the determinant factors, which is extremely important in cognitive load assessment because a wrong release of an individual can have bad consequences in some vital scenarios such as aerospace manipulation, surgical rescue, nuclear control and air traffic command. Moreover, since a single channel of physiological sensor data is often not faithful in cognitive

load assessment, e.g., a student may learn in a physical environment with high level of noise or high temperature, it is nonetheless difficult for these algorithms to specify only one kind of physiological signals like heart rate for desired load levels. This inspires us to recognize cognitive load states by discovering casual features from multivariate physiological sensor data.

To address these issues in cognitive load recognition, we present a hybrid spatio-temporal causal model by employing *Granger causality* (GC) and *Gramian angular difference fields* (GADF) to discover and combine multivariate physiological features. In particular, our approach considers a principled way of dealing with the inherent spatio-temporal causal variability within physiological signals. Briefly speaking, to discover causal structures in a single physiological channel such as heart rate, we present to introduce a temporal causal network (or *GC network*) generated from Granger causality test among physiological events. Now each resulting casual network contains its unique set of directed links together with their weights that represent cause-effect relations, characterizing a certain instance of a single channel that possess similar physiological features and their temporal causal dependencies. In addition, to combine the representative physiological features from multiple channels, we treat multivariate physiological signals as video-like continuous 2D objects (or called *GADF map*) by adopting GADF to characterize the inherit spatio-temporal dependency among different signals. Specifically, a capsule network is designed to merge the two heterogenous types of features, i.e., GC network and GADF map, together by leveraging the *encoder-classifier* mechanism to efficiently capture their spatio-temporal causal relations in a uniform way. In this way, our hybrid model is more capable of characterizing the inherit causal structural variability together with the spatio-temporal dependencies in cognitive load recognition when compared to existing methods, which is also verified during empirical evaluations on one publicly-available dataset and two in-house datasets under virtual reality (VR) environment collected by ourselves to be detailed in later sections.

2 Related Work

2.1 Knowledge-driven associations between physiological signals and cognitive load

There has been a fair amount of work on learning and recognizing cognitive load states by employing physiological signal data, much of them addressed from a “univariant” perspective. The first study can be traced back to 1963 when Kalsbeek [12] used ECG to analyze cognitive load. Nowadays, a variety of physiological events are studied to associate with the cognitive load states. For instance, different frequency bands in EEG spaces can achieve cognitive load discrimination within tasks [4, 25]. The ECG median absolute deviation and median heart flux are found to be the most accurate measurements at distinguishing levels of cognitive load [10]. HRV and PPG that reflect the states of heart activity and blood vessels behave different trends under different levels of cognitive load [27]. Other physiological signals such as galvanic skin response (GSR), respiration

(RESP) and electrodermal activity (EDA) have also been used as a measurement criterion for cognitive load assessment [2]. These approaches are capable of capturing rich relations, but unfortunately the semantic rules and their weights are typically hand-coded or based on domain knowledge. In particular, it is not practicable to handcraft the rules whose relations among physiological events are intricate especially for the multivariate signals.

2.2 Data-driven models for cognitive load assessment

Feature selection-based methods utilize features extracted from one or more specific signals to detect cognitive load. Most of these methods [22, 28] use traditional machine learning methods such as SVM and KNN as classifiers. Moreover, these methods need prior knowledge to decide which features are appropriate in cognitive load recognition. Currently, deep network-based approaches have been at the forefront of this research field. RNN [15] and LSTM [11] are widely implemented for cognitive load assessment, which are adopted to capture the temporal features. However, neither of them takes into account the spatio-temporal connections between physiological signals, and they are computationally expensive and difficult for parallel computing due to their sequential structures. To fully exploiting spatio-temporal dependencies, a series of CNN-based model and its variants such as FCN [24], MCDCNN [30], MCNN [6], CNN-LSTM [13] and MLSTM-FCN [13] are introduced to manage both spatial relationship from physiological signals. However, these approaches are limited to capture the spatio-temporal features from multiple physiological signals and ignore the causality among physiological events.

2.3 Granger causality

As aforementioned in the previous section, currently either knowledge-driven models or data-driven models are rather limited in further uncovering rich cause-effect relationships. Granger causality [9] is a way that can investigate causality between two physiological events that combines temporal relations with probabilistic description. GC-based model can capture event interactions and their temporal dependencies. Especially, it demonstrates the effectiveness in exploring causal event sets. In the field of cognitive load assessment existing GC-based models [18] exploit temporal dependencies between time series from raw physiological signals and use them to detect physiological events. However, they usually lack the expressive power to capture and propagate rich temporal dependencies in physiological events. Most importantly, since cause and effect are unidirectional, these models have to check triangle relationships to maintain causal consistency, which implies temporal consistency in the meantime. These methods often uses GC as a tool to discover temporal dependencies but fail to maintain causal consistency, which are computationally expensive or even intractable in discovering causal dependencies, where the event size is large. Moreover, it is difficult or even meaningless to understand the causes and effects that are learned from raw time series. It is worth clarifying that Granger causality does not imply

“true” causality since the question of “true causality” is deeply philosophical. It can be thought of as a tool of specifying a necessary condition for a temporal causal relation. To address the problems in these models, we present our hybrid model to explicitly capture the inherent causal structural varieties by combining physiological event-based causal networks together with spatio-temporal dependency map of multivariate physiological signals under consistency.

3 Problem Formulation

Given a dataset \mathcal{D} collected from C channels of physiological signals, a hybrid model is constructed with respect to the temporal causal relations as well as spatio-temporal maps among multivariate physiological events. Each sample is a sequence of T physiological events, denoted by $\mathbf{S} = \langle \mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_T \rangle$. A *physiological event* (or *event* for short) \mathbf{s}_t is a vector of C attributes at time interval t , with each being associated with a certain physiological channel. We denote it as $\mathbf{s}_t = (\mathbf{s}_{t1}, \mathbf{s}_{t2}, \dots, \mathbf{s}_{tC})$, where \mathbf{s}_{tc} is a vector of K data points collected from the c -th channel measured within the t -th time interval, written by $\mathbf{s}_{tc} = \langle s_{tc}(1), \dots, s_{tc}(K) \rangle$. In addition, a sequence of k ($k \leq K$) continuous observations in an individual channel event \mathbf{s}_{tc} is denoted by $\bar{\mathbf{s}}_{tc}(k) = \langle s_{tc}(1), \dots, s_{tc}(k) \rangle$. It is worth noting that all events are synchronized for any channel and are spaced at a uniform time interval of length K .

GC network. For each individual channel c , a GC network can be used to represent the temporal causal relationships between physiological events, where a node v_{tc} represents the corresponding physiological event \mathbf{s}_{tc} and a directed link describes the temporal causality between two related events. In what follows, we ignore the subscript c in the network for simplicity. Denote a GC network $\mathbf{X}_{gc} = (\mathbf{V}, \mathbf{E})$ the corresponding network of a sample \mathbf{S} , where \mathbf{V} is a set of T nodes. An event \mathbf{s}_i is a direct cause of \mathbf{s}_j if there is a directed link from v_i to v_j in \mathbf{E} , denoted by $v_i \rightarrow v_j$, where $v_i, v_j \in \mathbf{V}$. Any link $v_i \rightarrow v_j$ in a GC network \mathbf{X}_{gc} must satisfy *Granger causality test*, which defines v_i as the cause of v_j if the past values of v_i contain helpful information for predicting the future value of v_j . More formally, for each channel c , given the sequences of k observations of \mathbf{s}_{ic} and \mathbf{s}_{jc} ($k < K$), v_i is the cause of v_j with respect to data point k if $P(s_{jc}(k+1) \mid \bar{\mathbf{s}}_{ic}(k), \bar{\mathbf{s}}_{jc}(k)) \neq P(s_{jc}(k+1) \mid \bar{\mathbf{s}}_{jc}(k))$, and also states that v_i is not the cause of v_j if $P(s_{jc}(k+1) \mid \bar{\mathbf{s}}_{ic}(k), \bar{\mathbf{s}}_{jc}(k)) = P(s_{jc}(k+1) \mid \bar{\mathbf{s}}_{jc}(k))$. Since causality is transitive, irreflexive and anti-symmetric, it can be verified that the resulting GC network is a directed acyclic graph. A GC network should be *consistent* that the temporal causal relations on every triangle of nodes $\triangle ijk$ in the network satisfy the transitivity property such that if $v_i \rightarrow v_j$ and $v_j \rightarrow v_k$ then $v_k \nrightarrow v_i$. In this way, for each channel a network can characterize only a possible style (or an instance) of a cognitive load state.

GADF map. To capture the spatio-temporal correlation between physiological events, a unique feature map \mathbf{X}_{gadf} is generated for each event \mathbf{s}_{tc} . Here, each data point $s_{tc}(i)$ can be represented in polar coordinates by encoding its corresponding angular cosine value $\phi(i) = \arccos(s_{tc}(i))$ with the radius

$\rho(i) = \frac{i}{I}$, where I is a constant factor to regularize the span of the polar coordinate system. Due to the monotonicity of the cosine function in $[0, \pi]$, each channel of an event can be used to generate a unique polar map. Moreover, the temporal dependence between elements in a event can be preserved through the property of the varying radius $\rho(i)$. In this way, for any physiological event, we can readily identify spatial-temporal correlations by measuring the trigonometric differences between any pair of its corresponding points, i.e., a GADF map, defined as $\mathbf{X}_{gadf} = [\sin(\phi(i) - \phi(j))]_{i,j=1,\dots,T}$, which is a $T \times T$ matrix.

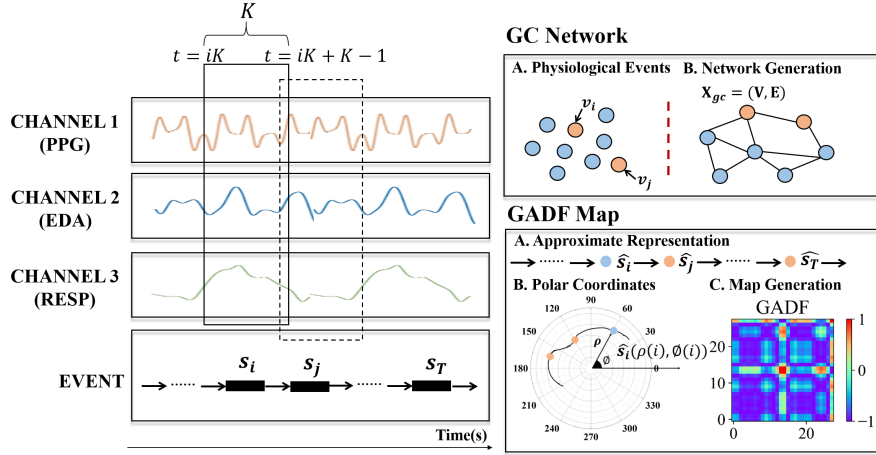


Fig. 1. Illustration of physiological events and their corresponding GC network and GADF map.

As shown in Fig. 1, GC network and GADF map can form a mixing feature space that describes a unique cognitive load states. This inspires us to present in what follows a hybrid model where these temporal causal and spatio-temporal features can be systematically discovered and combined to characterize the cognitive states of interests.

4 Our Approach

Let us consider a dataset \mathcal{D} of M samples $\{(\mathbf{S}_m, y_m)\}$ over Y classes (i.e. different levels of cognitive load states), where y_m is the label of the sample \mathbf{S}_m , $1 \leq m \leq M$. Here each sample $\mathbf{S}_m \in \mathcal{D}$ is associated with C -channel sequences of T physiological events $\mathbf{s}_t = \{s_{tc}\}_{c=1}^C$, $1 \leq t \leq T$. Our objective is to construct GC networks and GADF maps and encode them in a uniform way from these physiological events for cognitive load recognition tasks. The overview of our approach is illustrated in Fig. 2.

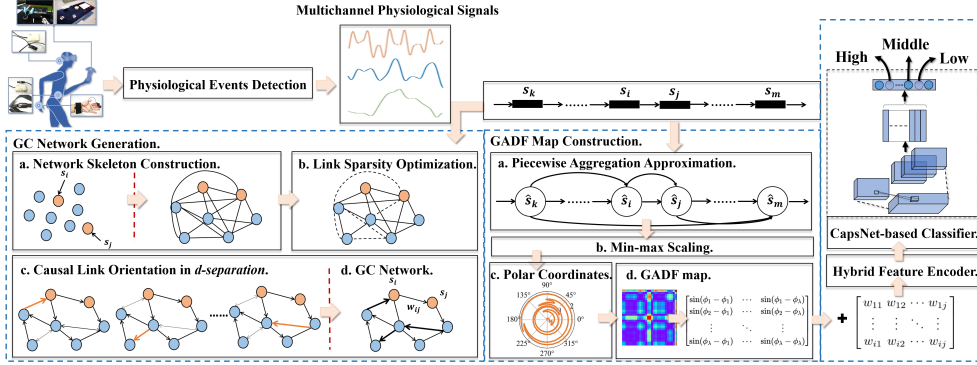


Fig. 2. The overall framework of our approach.

4.1 GC Network Generation

There are two steps to generate a GC network, i.e., network skeleton construction and causal link orientation.

Network Skeleton Construction. We first determine the network skeleton, i.e., which pairs of nodes (events) and their links (temporal causal relations) should be considered as candidates in the network. Formally, given two events of K observations of data points of c -th channel s_{ic} and s_{jc} , which are individually and jointly stationary, s_{jc} causes s_{ic} if adding s_{jc} helps predict s_{ic} , according to the definition of *Granger causality*. Subsequently, the jointly autoregressive model can be expressed as follows:

$$s_{ic}(k) = \sum_{\tau=1}^L b_{ii}(\tau) s_{ic}(k-\tau) + \sum_{\tau=1}^L b_{ij}(\tau) s_{jc}(k-\tau) + \beta_{ki}, \quad \beta_{ki} \sim \mathcal{N}(0, \Sigma_i), \quad (1)$$

$$s_{jc}(k) = \sum_{\tau=1}^L b_{jj}(\tau) s_{jc}(k-\tau) + \sum_{\tau=1}^L b_{ji}(\tau) s_{ic}(k-\tau) + \beta_{kj}, \quad \beta_{kj} \sim \mathcal{N}(0, \Sigma_j), \quad (2)$$

where $b_{ii}(\tau)$, $b_{jj}(\tau)$, $b_{ij}(\tau)$ and $b_{ji}(\tau)$ are regression coefficients, β_{ki} and β_{kj} are regression estimation residuals, and $\Sigma_i = \text{var}(\beta_{ki})$ and $\Sigma_j = \text{var}(\beta_{kj})$. L is a finite value called lag order, which can generally determined by Akaike Information Criterion (AIC).

More generally, for an individual channel c , we define the vector autoregression model regarding all pairs of physiological events (or nodes) as follows:

$$\mathbf{s}(k) = \sum_{\tau=1}^L \mathbf{B}(\tau) \mathbf{s}(k-\tau) + \beta_k, \quad (3)$$

where $\mathbf{B}(\tau)$ is the $T \times T$ coefficient matrix at lag τ where its entry $b_{ji}(\tau) \in \mathbf{B}(\tau)$ is the regression coefficient that indicates the effect on link $v_i \rightarrow v_j$, and β_k is

its corresponding residual vector of size T . We adopt the LASSO *algorithm* [1] to estimate these parameters as follows:

$$\begin{aligned}\hat{\mathbf{b}}_j &= \arg \min_{\mathbf{b}_j} \sum_{k=L+1}^K \|s_{jc}(k) - \sum_{i=1}^T \mathbf{b}_{ji}^\top \dot{\mathbf{s}}(k, L)\|_2^2 + \lambda \|\mathbf{b}_j\|_1 \\ &= \arg \min_{\mathbf{b}_j} \sum_{k=L+1}^K \|s_{jc}(k) - \sum_{i=1}^T \sum_{\tau=1}^L b_{ji}(\tau) s_{ic}(k - \tau)\|_2^2 + \lambda \|\mathbf{b}_j\|_1\end{aligned}\quad (4)$$

where \mathbf{b}_{ji} is the i -th vector of coefficients \mathbf{b}_j , i.e., $\mathbf{b}_{ji} = [b_{ji}(1), \dots, b_{ji}(L)]$, and $\dot{\mathbf{s}}(k, L)$ is the concatenated vector of L lagged observations, i.e. $\dot{\mathbf{s}}(k, L) = [s_{jc}(k-L), \dots, s_{jc}(k-1)]$. In this way, the links that have little influence between any pair of events (i.e., $b_{ji} \approx 0$) can be eliminated by the regularization in LASSO algorithm, and thereby ensuring the sparsity in the network, avoiding the exhaustive computation. Now we can construct the initial network skeleton \mathbf{X}_{gc}^* by setting $v_i \rightarrow v_j \in \mathbf{E}$ if and only if $\hat{\mathbf{b}}_{ji}$ is a nonzero vector.

Causal Link Orientation. Now there still exists the awkward situations where bidirectional links such as $v_i \leftrightarrow v_j$ or cyclic triangles (e.g., $v_i \rightarrow v_j \rightarrow v_k \rightarrow v_i$) exist in \mathbf{X}_{gc}^* , which may lead to causal inconsistency. To this end, we further orientate the links in \mathbf{X}_{gc}^* through the *d-separation* criterion, that is, if v_i and v_j are *d-separated* by v_k , then v_i and v_j are independent given v_k ; otherwise, v_i and v_j are interdependent given v_k . Here, we consider four types of *d-separation* based on the orientation rules [17]. After applying these rules, we can finally obtain a resulting GC network \mathbf{X}_{gc} that is causally inconsistent.

Besides, the weight on each link $v_i \rightarrow v_j$ can be estimated in terms of its causal power, as defined by:

$$w_{ij} = \begin{cases} \ln(\Phi_j/\Psi_{ij}), & \text{if } v_i \rightarrow v_j \in \mathbf{E} \text{ and } i \neq j \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

where Φ_j measures the prediction accuracy of v_j based on its own previous values, and Ψ_{ij} measures it from the previous values of both v_i and v_j . If $\Psi_{ij} < \Phi_j$, which means v_i have a causal influence on v_j . Theoretically, the larger w_{ij} , the stronger the causal influence.

4.2 GADF Map Construction

It is straightforward to construct a GADF map from an individual channel of an event \mathbf{s}_{tc} . Specifically, an approximate representation of \mathbf{s}_{ic} , written as \hat{s}_{tc} , can be calculated by applying a simple *piecewise aggregation approximation* [14], that is, $\hat{s}_{tc} = \frac{1}{K} \sum_{j=1}^K s_{tc}(j)$ ($t = 1, \dots, T$). Here, each \hat{s}_{tc} is normalized within the range of $[-1, 1]$. Next, we transform each event representation \hat{s}_{tc} to a pair $(\phi(t), \rho(t))$ in the polar coordinate system. Formally, a GADF map \mathbf{X}_{gadf} is a $T \times T$ matrix with its entry being calculated as:

$$\mathbf{X}_{gadf}(i, j) = \sin(\phi(i) - \phi(j)) = (\hat{s}_{ic} - \hat{s}_{jc}) \sqrt{1 - \hat{s}_{ic}^2} \sqrt{1 - \hat{s}_{jc}^2}. \quad (6)$$

It is verified that GADF maps can provide intuitive spatio-temporal details as well as a cross-boundary division [23].

4.3 Capsule Network-Based Recognition Model

Now we are ready to build a hybrid model that can merge these two types of encoded features (i.e., \mathbf{X}_{gc} and \mathbf{X}_{gadf}) together as new inputs for cognitive load state recognition. Here we design an encoder-classifier model, which consists of two parts: a hybrid feature encoder that discovers the deep features by combining GC network and GADF maps, and a capsule network-based classifier to achieve the tasks of classifying different levels of cognitive load states [29].

Hybrid Feature Encoder. The input feature tensor \mathbf{X} is a concatenation of $\mathbf{X}_{gc}, \mathbf{X}_{gadf} \in \mathbb{R}^{C \times T \times T}$ of all the channels, and thus $\mathbf{X} \in \mathbb{R}^{C \times 2 \times T \times T}$. First, a convolution layer F_{conv} aims to transform these causal and spatio-temporal information jointly into a higher-level feature space, where the output feature tensor is denoted by $\mathbf{Z} \in \mathbb{R}^{C' \times 2 \times T \times T}$ ($C' < C$), as defined:

$$\begin{aligned} \text{Layer ①: } F_{conv} : \mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_C) &\mapsto \mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_{C'}) \\ \text{with } \mathbf{z}_{c'} = \kappa_{c'} * \mathbf{X} &= \sum_{c=1}^C \kappa_{c'} * \mathbf{x}_c, c' = 1, \dots, C', \end{aligned} \quad (7)$$

where $\kappa_{c'}$ is a filter kernel and $*$ is the convolution operator.

Next, we compress the global spatial information from several separate channels by adopting the global average pooling (*gap*) layer, and its output is fed into two fully-connected (*fc*) layers with ReLU activation function and sigmoid function σ , as formulated:

$$\text{Layers ②-④: } \boldsymbol{\omega} = F_{fc}^2(F_{gap}(\mathbf{Z})) = \sigma(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot (\text{avg}(\mathbf{z}_c))_{c=1, \dots, C'})) \quad (8)$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{C' \times C'}$ are the corresponding weights.

The last layer of our encoder is defined by a *channel-wise soft-threshold operation*:

$$\begin{aligned} \text{Layer ⑤: } \mathbf{M} &= \mathbf{X} + \mathbf{Z} \downarrow \boldsymbol{\tau} \\ \text{with } \boldsymbol{\tau} &= \boldsymbol{\omega} \odot F_{gap}(\mathbf{Z}), \mathbf{Z} \downarrow \boldsymbol{\tau} = (\mathbf{z}_c \downarrow \tau_c)_{1 \leq c \leq C'} \end{aligned} \quad (9)$$

where \odot is the element-wise product, and \downarrow is the soft-threshold operation. In this way, $\boldsymbol{\omega}$ and $\boldsymbol{\tau}$ contain the scaling weights and the thresholds for all the channels, respectively.

CapsNet-Based Classifier. The capsule network is used as a classifier of cognitive load levels, where it takes the previous encoder's output $\mathbf{M} \in \mathbb{R}^{C' \times T \times T}$ as its input and output a vector of size Y indicating the different levels of cognitive load states. Our classifier consists of three layers: a standard convolutional layer, a primary capsule layer and a cognitive capsule layer.

In details, the standard convolutional layer has 64 different 3×3 filters with a stride of 2 and a ReLU activation function. The primary capsule layer has 64 types of primary capsules \mathbf{U}_i ($i = 1, \dots, 64$). Each \mathbf{U}_i is generated by a convolutional operation with 8 different 2×2 filters and then is reshaped as a tensor of 8D vectors $\tilde{\mathbf{U}}_i = [\tilde{\mathbf{u}}_{i,1}, \dots, \tilde{\mathbf{u}}_{i,d}]$ where $d = 8K^2$. Last, the cognitive capsule layer transforms these vectors to Y different 16D vectors by employing a specific *weighting* and *routing* procedure as follows:

$$\mathbf{F}_{wr} : \tilde{\mathbf{U}} = (\tilde{\mathbf{u}}_{i,k})_{i=1,\dots,64,k=1,\dots,d} \mapsto \mathbf{Y} = (\mathbf{y}_j)_{j=1,\dots,Y} \quad (10)$$

More specifically, \mathbf{F}_{wr} includes two steps. First, for each i ($i = 1, \dots, 64$), the primary capsules in $\tilde{\mathbf{U}}_i = (\tilde{\mathbf{u}}_{i,1}, \dots, \tilde{\mathbf{u}}_{i,d})$ pass through a shared 8×16 weight matrix $\mathbf{W}_{i,j}$ to generate $\hat{\mathbf{U}}_{j|i} = (\hat{\mathbf{u}}_{j|i,1}, \dots, \hat{\mathbf{u}}_{j|i,d})$ ($j = 1, \dots, Y$). Next, a *dynamic routing procedure* routes each primary capsule output $\hat{\mathbf{u}}_{j|i,k}$ to the j -th cognitive capsule and produces the output \mathbf{y}_j for all $i = 1, \dots, 64$ and $k = 1, \dots, d$. A squashing function is employed to ensure that short vectors get shrunk to almost zero length while long vectors get shrunk to almost a unit length, as defined as follows:

$$\mathbf{y}_j = \text{squash}(\mathbf{e}_j) = \frac{\|\mathbf{e}_j\|^2}{1 + \|\mathbf{e}_j\|^2} \times \frac{\mathbf{e}_j}{\|\mathbf{e}_j\|}, \quad (11)$$

where $\mathbf{e}_j = \sum_{i,k} \frac{\exp(q_{j|i,k})}{\sum_{j=1}^n \exp(q_{j|i,n})} \cdot \hat{\mathbf{u}}_{j|i,k}$ and $q_{j|i,k}$ is an internal parameter which is updated by $q_{j|i,k} \leftarrow q_{j|i,k} + \hat{\mathbf{u}}_{j|i,k} \cdot \mathbf{y}_j$ at each iteration. The loss function of our CapsNet-based classifier is defined below:

$$\text{Loss}_j = \mathbf{I}_j \max(0, m^+ - \|\mathbf{y}_j\|)^2 + \gamma (1 - \mathbf{I}_j) \max(0, \|\mathbf{y}_j\| - m^-)^2 \quad (12)$$

where \mathbf{I}_j is an indicator function that indicates whether the true label of a sample is class j , m^+ (*resp.* m^-) refers to the upper (*resp.* lower) boundary, and γ is a regularization weight. $\|\mathbf{y}_j\| \in [0, 1]$ and $\hat{j} = \max_j \{\|\mathbf{y}_j\|\}$ indicates the final result is recognized as the class of \hat{j} .

5 Empirical Evaluations

5.1 Datasets and Preprocessing

Three cognitive load assessment datasets are considered in our experiments, including one publicly-available cognitive load datasets and two in-house dataset on VR learning environment collected by ourselves.

CLAS [16]: This is a publicly-available dataset, which contains synchronized ECG, PPG, and EDA signals (256 Hz) captured from 62 subjects with each 30-minute recording involved in purposely designed interactive or perceptive task indicating two cognitive load states. According to the description of related paper, when the subjects were in the sub-task session, the cognitive load was high, while in the neutral stimulus session, the cognitive load was low. For a better

comparison with our data set, we set the original CLAS dataset as a sample set with a sliding window size of 5s.

3s-COGSET and **5s-COGSET**: To our best knowledge, the above mentioned dataset is so far the only one publicly available and suitable for deep learning methods in the field of cognitive load assessment. In particular, the instances of the cognitive tasks in the experiments of CLAS are relatively simple without considering the practicality of the test scenario. To this end, we conducted a new experiment, which is still an ongoing effort, and at the moment 16 subjects (8 male and 8 are female) with their ages ranging from 18 to 24 were recruited to learn 50 modules of courses that are designed by ourselves in VR environment. Each module is performed 10 runs by each participant. Three types of physiological signals, i.e., PPG, RESP and EDA, were recorded during performing the tasks by means of wearable sensors with the sampling rate of 64Hz. Our experiment contains around 5,000 annotated samples about three levels of cognitive load states (i.e., low, medium and high.) on VR learning environment. A subset of samples are provided in the supplementary material, and once ready we plan to share the entire dataset in the community. Considering the different settings of physiological events, our records were divided into two new datasets by using the event sizes K of 3s and 5s, respectively.

5.2 Experimental Set-Ups

Our model is implemented by Keras with backend of Tensorflow. It is optimized by Adam optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) with the learning rate of 1×10^{-4} and the step size of $e^{-0.1}$ on one GeForce GTX 750Ti GPU. We set the parameter $K = 28$, $m^+ = 0.9$, $m^- = 0.1$, $\gamma = 0.5$. The batch size is fixed to 14. We compare the classification performance of our model with 7 conventional models and 11 deep models. To make a fair comparison, we did not use any data augmentation or pre-trained weights to improve performance. The ratio of training and testing sequences is 4 : 1. Accuracy was used as the evaluation metric, which is calculated as the proportion of true results among the total number of samples.

5.3 Experimental Results

Comparison Against Conventional Models. Table 1 depicts the comparison results under different settings of physiological channels. In order to integrate various existing feature extraction methods, we extracted 787 features (e.g. Fast Fourier Transformation coefficients, etc.) in total, which however need to be manually encoded from prior knowledge. Generally, our model outperforms these models by a large margin. This is because our hybrid model is capable of capturing causal dependencies and spatio-temporal features among multivariate physiological events.

Comparison Against Other Deep Models. Table. 2 shows the comparison results with other deep models that recognize cognitive load states directly from

Table 1. Accuracy comparisons on three datasets under different settings of physiological channels.

Conventional models	Accuracy							
	LR	SVM	GNB	DT	RF	XGBoost	KNN	Ours
CLAS	0.60	0.61	0.59	0.55	0.60	0.64	0.66	0.75
3s-COGSET	0.49	0.58	0.57	0.55	0.67	0.63	0.70	0.86
5s-COGSET	0.58	0.65	0.51	0.61	0.54	0.75	0.77	0.92
under different combinations of physiological signals								
5s-COGSET (PPG)	0.49	0.49	0.54	0.51	0.48	0.54	0.57	0.70
5s-COGSET (RESP)	0.48	0.50	0.54	0.56	0.55	0.60	0.59	0.65
5s-COGSET (EDA)	0.49	0.50	0.49	0.48	0.51	0.55	0.58	0.62
5s-COGSET (PPG+EDA)	0.60	0.54	0.54	0.46	0.63	0.70	0.69	0.79
5s-COGSET (PPG+RESP)	0.60	0.60	0.62	0.66	0.65	0.64	0.70	0.76
5s-COGSET (RESP+EDA)	0.62	0.60	0.64	0.59	0.67	0.66	0.74	0.81

raw physiological signals. Apparently, it can be observed that our model can is significantly more accurate than other models with around 5%-30% performance boost. Notably, MLP and MCDCNN get relatively acceptable results of identifying states. This is mainly due to their abilities to take advantage of the rich hierarchical and temporal dependency information between various physiological events. It is also clear that our model is superior to other models including those that combine CNN and LSTM (or RNN) structures that can also capture spatio-temporal dependencies among multivariate signals. This is mainly due to the reason that GC network can describe the temporal causal relation between any pair of events.

Table 2. Accuracy comparisons against other deep models. The percentage in the bracket shows the accuracy change taken our approach as a baseline.

Deep models	Accuracy		
	CLAS	3s-COGSET	5s-COGSET
MLP [24]	0.67(-0.08)	0.80(-0.06)	0.85(-0.07)
FCN [24]	0.69(-0.06)	0.57(-0.29)	0.58(-0.34)
ResNet [24]	0.70(-0.05)	0.61(-0.25)	0.56(-0.36)
Inception [8]	0.61(-0.14)	0.67(-0.19)	0.70(-0.22)
MCDCNN [30]	0.60(-0.15)	0.76(-0.10)	0.84(-0.08)
MCNN [6]	0.54(-0.21)	0.57(-0.29)	0.56(-0.36)
1D-CapsNet [3]	0.50(-0.25)	0.75(-0.11)	0.60(-0.32)
Parallel CNN-LSTM [13]	0.63(-0.12)	0.76(-0.10)	0.83(-0.09)
Serial CNN-LSTM [19]	0.56(-0.19)	0.50(-0.46)	0.49(-0.43)
MLSTM-FCN [13]	0.61(-0.14)	0.57(-0.29)	0.61(-0.31)
Grid-CNNs [26]	0.45(-0.30)	0.49(-0.37)	0.56(-0.36)
Ours	0.75	0.86	0.92

Convergence Speed Fig. 3(a) displays the training time of our model. It can be seen that our model converges after 30 epochs. Fig. 3(b) reports the comparison results of convergence speeds among different models. Notably, our model converges faster than other methods, which is beneficial to the training and optimization process. Theoretically, the time complexity of our models consists of three parts $O(MTK^2)$, $O(MTK^2)$ and $O(\sum_{l=1}^H M_l^2 K_l^2 H_{l-1} H_l)$, indicating the GC network generation, GADF map construction and capsule network-based classifier, respectively. H represents the number of layers of the classifier, and H_{l-1} and H_l refer to the sizes of input and output feature tensors at the l -th layer, respectively.

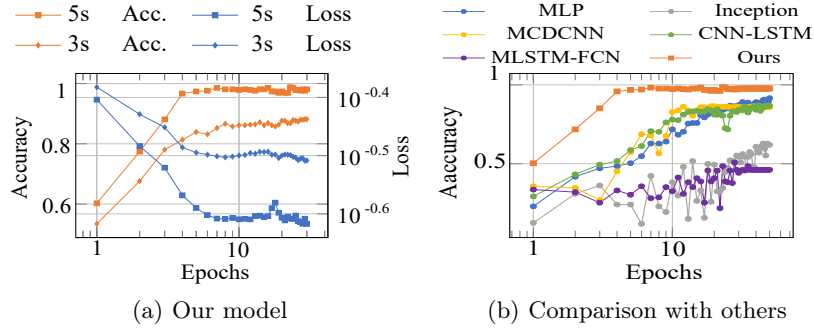


Fig. 3. Convergence speed comparison.

5.4 Ablation Study

In this section, we conduct three ablation studies to measure the effectiveness of the modules in our model.

Feature Encoding. We compared our hybrid features of GC network and GADF map with other three commonly used encoded features, i.e., Markov Transition Fields (MTF) [23], Recurrence Plot(RP) [26] and a simple grid structure (Grid) [7]. Fig. 4(a) shows that our hybrid features clearly outperform other encoded features in accuracy on the two in-house datasets. This is because the hybrid features contain not only the temporal causal configurations of a particular cognitive state but also the inherit spatio-temporal dependency among multivariate signals.

Optimum Parameter Selection. We also compared various settings of lag order L in our model. Here we increase the lag order L from 1 to 10 with a step of 1. The result shows that changing the lag order cannot lead to negative effects

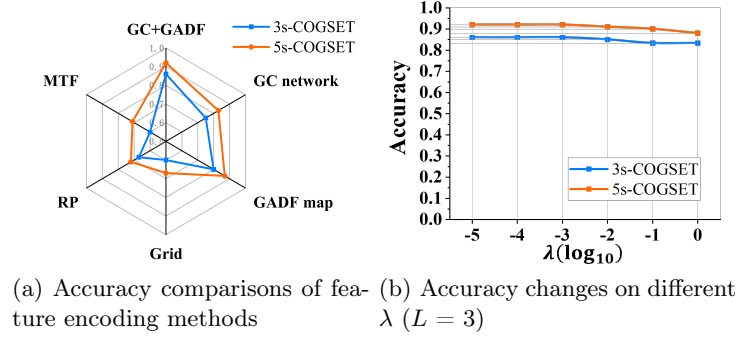


Fig. 4. Convergence speed comparison.

on the performance of our model on the datasets. This is mainly because the duration of cognitive load responses in a subsequence is very short. For instance, there are instantaneous loads, which fluctuate every moment from the beginning to the end of performing a task or set of tasks, such as cognitive dissonance and cognitive overload in a certain subsequence, so the variation of L is limited to affect the final results. Although the selection of lag order still remains an open issue, we suggest to set the lag order with a value that is slightly larger than the ordinary length of physiological events. Note that a very large value of L may result in computational burden.

The sparsity regularization parameter λ in Eq.(4) is an important parameter for link sparsity optimization. Its effect on classification performance on the three datasets is shown in Fig. 4(b) by fixing the lag order to $L = 3$. It is clear that increasing the value of λ strengthens the regularization effect. On the other hand, a small value of λ will bring about a great number of noisy links in the network, which may also be unfavorable to the recognition results.

Encoder-Classifier Component Effectiveness. The effectiveness of different components in encoder-classifier mechanism are separately evaluated by removing or replacing them with other conventional models. We evaluated two types of modules, including the encoder (i.e., remove the encoder and directly use the raw GC network and GADF map as input) and the classifier (i.e. remove the CapsNet-based classifier and only adopt a one-dense-layer for classification). Table. 3 reports the comparison results on the two in-house datasets, which indicates that changing the components may have a negative impact on the performance of our model. Obviously, classification performance degrades when either component is removed. This might be due to the hybrid encoding of both causal and spatio-temporal information in a uniform way in our model. Besides, when removing both components, the model gives worse performance than that using either our encoder or classifier, which indicates that our model is more

effective to capture causal and spatio-temporal dependencies at the same time than obtaining either of them individually.

Table 3. The impact of the components in our model. \times means no such component, while \checkmark denotes the reservation of it.

No.	Encoder	Classifier	Accuracy	
			3s-COGSET	5s-COGSET
1	\times	\times	0.54	0.52
2	\checkmark	\times	0.55	0.53
3	\times	\checkmark	0.76	0.85
4	\checkmark	\checkmark	0.86	0.92

6 Conclusion and Future Work

In this paper, we present a hybrid cognitive load recognition model by merging Granger causality network and Gramian angular difference fields map together for multivariate physiological data, which can capture the inherit causal and spatio-temporal varieties of physiological events in a uniform way. It is more efficient and flexible than existing methods on cognitive load recognition. As for future work, we will explore the applications of our model on more VR learning classes, and we will consider extending our model to detect multiple cognitive states with probabilities and will instead learn a model under uncertainty.

Acknowledgement This work was supported by grants from the National Major Science and Technology Projects of China (grant no. 2018AAA0100703), the National Natural Science Foundation of China (grant nos. 61977012, 61977054), the Central Universities in China (grant no. 2021CDJYGRH011).

References

1. Arnold, A., Liu, Y., Abe, N.: Temporal causal modeling with graphical granger methods. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 66–75 (2007)
2. Barua, S., Ahmed, M.U., Begum, S.: Towards intelligent data analytics: A case study in driver cognitive load classification. Brain sciences **10**(8), 526 (2020)
3. Butun, E., Yildirim, O., Talo, M., Tan, R.S., Acharya, U.R.: 1d-cadcapsnet: One dimensional deep capsule networks for coronary artery disease detection using ecg signals. Physica Medica **70**, 39–48 (2020)
4. Chakladar, D.D., Dey, S., Roy, P.P., Dogra, D.P.: Eeg-based mental workload estimation using deep blstm-lstm network and evolutionary algorithm. Biomedical Signal Processing and Control **60**, 101989 (2020)

5. Critchley, H.D., Garfinkel, S.N.: The influence of physiological signals on cognition. *Current Opinion in Behavioral Sciences* **19**, 13–18 (2018)
6. Cui, Z., Chen, W., Chen, Y.: Multi-scale convolutional neural networks for time series classification. arXiv preprint arXiv:1603.06995 (2016)
7. Eckmann, J.P., Kamphorst, S.O., Ruelle, D., et al.: Recurrence plots of dynamical systems. *World Scientific Series on Nonlinear Science Series A* **16**, 441–446 (1995)
8. Fawaz, H.I., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D.F., Weber, J., Webb, G.I., Idoumghar, L., Muller, P.A., Petitjean, F.: Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery* **34**(6), 1936–1962 (2020)
9. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**(3), 424–438 (1969)
10. Haapalainen, E., Kim, S., Forlizzi, J.F., Dey, A.K.: Psycho-physiological measures for assessing cognitive load. In: *Proceedings of the 12th ACM international conference on Ubiquitous computing*. pp. 301–310 (2010)
11. Hefron, R.G., Borghetti, B.J., Christensen, J.C., Kabbani, C.M.S.: Deep long short-term memory structures model temporal dependencies improving cognitive workload estimation. *Pattern Recognition Letters* **94**, 96–104 (2017)
12. Kalsbeek, J., Ettema, J.: Continuous recording of heart rate and the measurement of perceptual load. *Ergonomics* **6**(3), 306–307 (1963)
13. Karim, F., Majumdar, S., Darabi, H., Harford, S.: Multivariate lstm-fcns for time series classification. *Neural Networks* **116**, 237–245 (2019)
14. Keogh, E.J., Pazzani, M.J.: Scaling up dynamic time warping for datamining applications. In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 285–289 (2000)
15. Kuanar, S., Athitsos, V., Pradhan, N., Mishra, A., Rao, K.R.: Cognitive analysis of working memory load from eeg, by a deep recurrent neural network. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 2576–2580. IEEE (2018)
16. Markova, V., Ganchev, T., Kalinkov, K.: Clas: A database for cognitive load, affect and stress recognition. In: *2019 International Conference on Biomedical Innovations and Applications (BIA)*. pp. 1–4. IEEE (2019)
17. Meek, C.: Causal inference and causal explanation with background knowledge. arXiv preprint arXiv:1302.4972 (2013)
18. Ning, Y., Li, K., Zhang, Y., Chen, P., Yin, D., Zhu, H., Jia, H.: Assessing cognitive abilities of patients with shift work disorder: insights from rbans and granger causality connections among resting-state networks. *Frontiers in Psychiatry* p. 780 (2020)
19. Oord, A.v.d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
20. Rim, B., Sung, N.J., Min, S., Hong, M.: Deep learning in physiological signal data: A survey. *Sensors* **20**(4), 969 (2020)
21. Sweller, J.: Cognitive load during problem solving: Effects on learning. *Cognitive science* **12**(2), 257–285 (1988)
22. Wang, C., Guo, J.: A data-driven framework for learners' cognitive load detection using ecg-ppg physiological feature fusion and xgboost classification. *Procedia computer science* **147**, 338–348 (2019)
23. Wang, Z., Oates, T.: Imaging time-series to improve classification and imputation. In: *Twenty-Fourth International Joint Conference on Artificial Intelligence* (2015)

24. Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: A strong baseline. In: 2017 International joint conference on neural networks (IJCNN). pp. 1578–1585. IEEE (2017)
25. Xiong, R., Kong, F., Yang, X., Liu, G., Wen, W.: Pattern recognition of cognitive load using eeg and ecg signals. *Sensors* **20**(18), 5122 (2020)
26. Ye, Y., Jiang, J., Ge, B., Dou, Y., Yang, K.: Similarity measures for time series data classification using grid representation and matrix distance. *Knowledge and Information Systems* **60**(2), 1105–1134 (2019)
27. Yu, J., Liu, G.Y., Wen, W.H., Chen, C.W.: Evaluating cognitive task result through heart rate pattern analysis. *Healthcare Technology Letters* **7**(2), 41–44 (2020)
28. Zhang, X., Lyu, Y., Qu, T., Qiu, P., Luo, X., Zhang, J., Fan, S., Shi, Y.: Photoplethysmogram-based cognitive load assessment using multi-feature fusion model. *ACM Transactions on Applied Perception (TAP)* **16**(4), 1–17 (2019)
29. Zhao, M., Zhong, S., Fu, X., Tang, B., Pecht, M.: Deep residual shrinkage networks for fault diagnosis. *IEEE Transactions on Industrial Informatics* **16**(7), 4681–4690 (2019)
30. Zheng, Y., Liu, Q., Chen, E., Ge, Y., Zhao, J.L.: Time series classification using multi-channels deep convolutional neural networks. In: International conference on web-age information management. pp. 298–310. Springer (2014)