MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Universal Physiological Representation Learning with Soft-Disentangled Rateless Autoencoders

Han, Mo; Ozdenizci, Ozan; Koike-Akino, Toshiaki; Wang, Ye; Erdogmus, Deniz

TR2021-027 April 06, 2021

Abstract

Human computer interaction (HCI) involves a multidisciplinary fusion of technologies, through which the control of external devices could be achieved by monitoring physiological status of users. However, physiological biosignals often vary across users and recording sessions due to unstable physical/mental conditions and taskirrelevant activities. To deal with this challenge, we propose a method of adversarial feature encoding with the concept of a Rateless Autoencoder (RAE), in order to exploit disentangled, nuisance-robust, and universal representations. We achieve a good trade-off between user-specific and task-relevant features by making use of the stochastic disentanglement of the latent representations by adopting additional adversarial networks. The proposed model is applicable to a wider range of unknown users and tasks as well as different classifiers. Results on cross-subject transfer evaluations show the advantages of the proposed framework, with up to an 11.6% improvement in the average subject-transfer classification accuracy.

IEEE Journal of Biomedical and Health Informatics

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Universal Physiological Representation Learning with Soft-Disentangled Rateless Autoencoders

Mo Han, Ozan Özdenizci, *Member, IEEE*, Toshiaki Koike-Akino, *Senior Member, IEEE*, Ye Wang, *Senior Member, IEEE*, and Deniz Erdoğmuş, *Senior Member, IEEE*

***Abstract*— Human computer interaction (HCI) involves a multidisciplinary fusion of technologies, through which the control of external devices could be achieved by monitoring physiological status of users. However, physiological biosignals often vary across users and recording sessions due to unstable physical/mental conditions and task-irrelevant activities. To deal with this challenge, we propose a method of adversarial feature encoding with the concept of a Rateless Autoencoder (RAE), in order to exploit disentangled, nuisance-robust, and universal representations. We achieve a good trade-off between user-specific and task-relevant features by making use of the stochastic disentanglement of the latent representations by adopting additional adversarial networks. The proposed model is applicable to a wider range of unknown users and tasks as well as different classifiers. Results on cross-subject transfer evaluations show the advantages of the proposed framework, with up to an 11.6% improvement in the average subject-transfer classification accuracy.**

***Index Terms*— stochastic bottleneck, soft disentanglement, disentangled representation, deep learning, autoencoders, adversarial learning, physiological biosignals**

I. Introduction

HUMAN computer interaction (HCI) [1] is a fundamental technology enabling machines to monitor physiological disorders, to comprehend human emotions, and to execute proper actions, so that users can control external devices through their physiological status in a safe and reliable fashion. To measure traditional physiological biosignals such as electrocardiogram (ECG) [2], electromyography (EMG) [3], [4] and electroencephalography (EEG) [5–8], either implanted or surface electrodes and their frequent calibration are necessary, reducing user comfort while increasing the overall expense. Recently, novel wearable sensors such as wrist-worn devices were developed for accurately measuring physiological signals [9–13] (e.g., arterial oxygen level, heart rate, skin temperature, etc.) in comfortable and effective manners. Utilizing these non-EEG physiological biosignals can effectively increase the system convenience during data collection with less expense.

One major challenge of physiological status assessment lies in reducing the variability in biosignals across users or recording sessions due to the unstable mental/physical conditions and task-irrelevant disturbances. Addressing biosignal datasets collected from a narrow amount of subjects, transfer learning methods [14–17] are applied to build strong feature learning machines to extract robust and invariant features across various tasks and/or unknown subjects. Particularly, adversarial transfer learning [18–24] demonstrated impressive results in constructing such discriminative feature extractors.

Traditional adversarial transfer learning works [20–24] aim to extract latent representations universally shared by a group of attributes using adversarial inference, where a discriminative network is trained adversarially towards the feature extractor in order to differentiate universal features from various attributes. For example, in [21], the authors use an adversarial method to learn modality-shared features for cross-modal retrieval, with an adversarial discriminative model distinguishing the modalities to the generative model. Similarly, Sun et al. [22] explore cross-project defect representations by adversarially training the feature transformer and project discriminator. In [23] an adversarial classifier was also trained when attached to the feature encoder for making subject-invariant decisions. However, in those existing approaches, the adversarial training scheme is usually applied indiscriminately on the whole feature group with only one discriminative network, which inevitably leads to the loss of attribute-discriminative information. Therefore, rather than using only one adversarial discriminator to merely preserve shared cross-attribute features, we train two additional adversarial discriminators jointly with the feature extractor, so that the physiological features could be disentangled into two counterparts representative of subject and task associated information respectively. In this way, the variability in both subject and task space can be better accounted for.

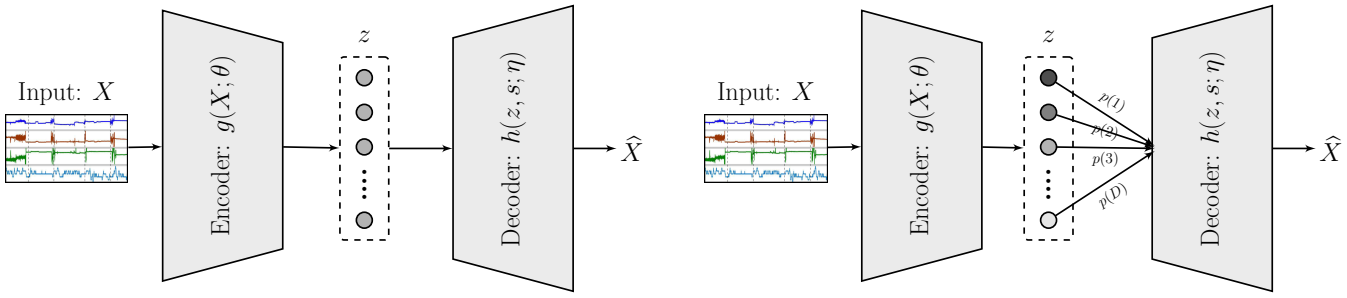
As a commonly used feature extractor framework for transfer learning, autoencoders (AE) [25–27] can learn latent representations with a dimensionality typically much smaller than the input data, which is known as a "bottleneck" architecture, while capturing key data features to enable data reconstruction from the latent representation. A challenging problem in

M. Han and D. Erdoğmuş are with Cognitive Systems Laboratory, Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, USA. E-mail: {han, erdogmus}@ece.neu.edu

O. Özdenizci is with the Institute of Theoretical Computer Science, Graz University of Technology, Austria. E-mail: oezdenizci@tugraz.at

T. Koike-Akino and Y. Wang are with Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA. E-mail: {koike, yewang}@merl.com

M. Han was an intern at MERL during this work. D. Erdoğmuş is partially supported by NSF (IIS-1149570, CNS-1544895, IIS-1715858), DHHS (90RE5017-02-01), and NIH (R01DC009834).



(a) Conditional autoencoder (cAE)

(b) Conditional rateless autoencoder (cRAE)

Fig. 1: (a) Conditional autoencoder (cAE): an encoder-decoder pair where the encoder estimates latent $z = g(X;\theta)$ with parameters θ, and the decoder estimates reconstructed input signals $\hat{X} = h(z,s;\eta)$ with parameters η, using the latent z and conditioning variable s. When decoder is $h(z;\eta)$, it reduces to a traditional autoencoder (AE). (b) Conditional rateless autoencoder (cRAE): a probabilistic cAE model with a stochastic bottleneck where dth latent representation node is assigned with dropout probability rates $p(d)$, such that the conditional decoder takes a subset of the latent units as input.

dimensionality reduction is to determine an optimal feature dimensionality which sufficiently captures latent information that is essential for particular tasks. To address this issue, the Rateless Autoencoder (RAE) [28] was proposed to enable the AE to seamlessly adjust feature dimensionality through its rateless property, while not requiring a fixed structure of bottleneck. To realize such flexibility in the latent space, RAE implements a probabilistic latent dimensionality which is stochastically decreased through dropout during training, where a non-uniform dropout rate distribution is imposed to the bottleneck structure.

In this work, we propose an adversarial learning method with the RAE concept newly introduced, extended from [18] and [19]. In [18] and [19], the entire latent feature group were hard-split into task-related and subject-relevant parts and attached to two adversarial classifiers respectively, where the selection of split ratio is challenging when the underlying nature of the bottleneck is still vague, and the trade-off and smooth transition in between task- relevant and irrelevant counterparts could be lost. In the proposed method, we exploit rateless soft-disentangled representations by connecting all features to two discriminators with different dropout rates instead of a hard split. More specifically, existing adversarial methods [18–24] are all special cases of the proposed method with different dropout rate distributions when connecting adversarial classifiers to the encoder. Our contributions are summarized as follows:

- We complementarily use two additional adversarial networks, i.e., adversary and nuisance blocks, to disentangle and re-organize the latent representations.
- The rateless trade-off between subject-specific and task-relevant features is exploited by stochastically attaching adversary and nuisance blocks to the encoder.
- Different dropout strategies of the disentangled adversarial RAE are discussed.
- Empirical assessments were performed on a publicly available dataset of physiological biosignals for measuring human stress level through cross-subject evaluations with various classifiers.
- Comparative experiments on multiple models including traditional AE and adversarial methods are evaluated.
- We demonstrate the remarkable advantage of the proposed framework, achieving up to an 11.6% improvement in subject-transfer classification accuracy.

II. Methodology

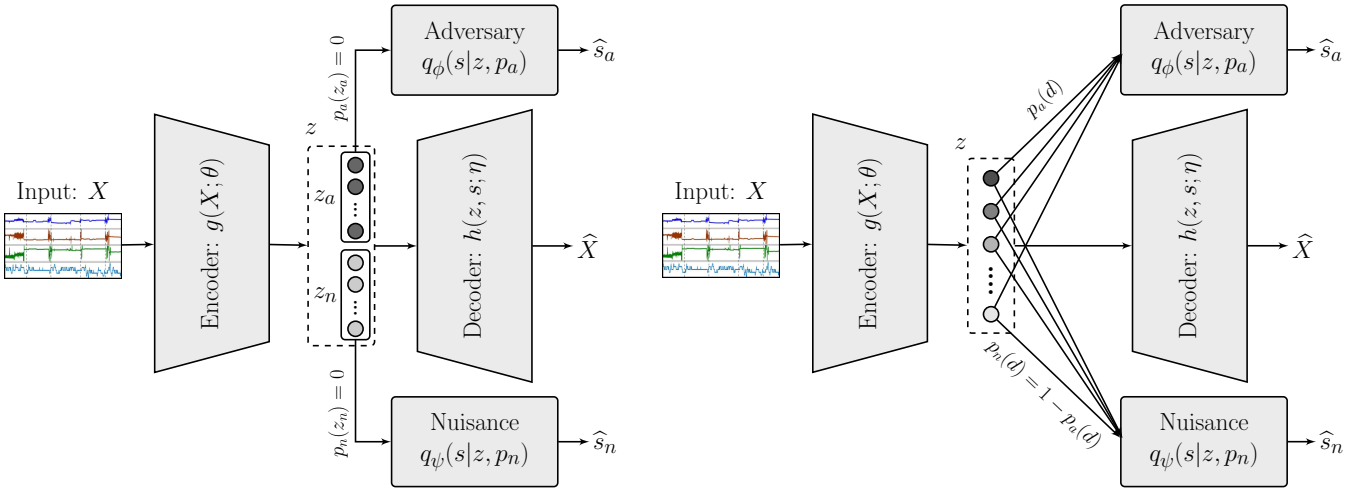
A. Notation and Problem Description

We define $\{(X_i, y_i, s_i)\}_{i=1}^{n}$ as a labeled data set, where $X_i \in \mathbb{R}^C$ is the input data vector recorded from C channels of trial i, $y_i \in \{0, 1, \ldots, L-1\}$ is the class label of user task/status among L classes, and $s_i \in \{1, 2, \ldots, S\}$ is the user identification (ID) index among S subjects. The task/status y is assumed to be marginally independent with respect to subject ID s, and the physiological signal is generated dependently on y and s, i.e., $X \sim p(X|y,s)$. The aim is to construct a model to estimate the task/status label y given an observation X, where the model is generalized across the variability of subject s, which is considered as a nuisance variable associated with transferring the feature extraction model.

B. Rateless Autoencoder (RAE)

AE is a well-known feature learning machine which includes a network pair of encoder and decoder, as shown in Fig. 1(a). The encoder packs data features into a latent representation z, while the decoder intends to re-construct the input data X based on the latent representation z. AE structures are typically bottleneck architectures, where the dimensionality D of representation z is lower than the dimensionality of input data X, and the latent variables should contain adequate features capable of reconstructing the original data through its corresponding decoder network. A challenging problem in such a dimensionality reduction is to decide an optimal feature dimensionality which captures sufficient latent representations that are essential for specific tasks.

RAE [28] is an AE family providing a rateless property that enables the AE to seamlessly adjust feature dimensionality.



(a) Disentangled adversarial autoencoder (DA-cAE)

(b) Disentangled adversarial rateless autoencoder (DA-cRAE)

Fig. 2: Disentanlgled adversarial autoencoder for nuisance-robust transfer learning. (a) Disentangled adversarial conditional autoencoder (DA-cAE) with hard split: a deterministic disentangled universal latent representation learning model where z is partitioned into sub-parts z_a and z_n which are adversarially trained to be s-invariant (i.e., z_a used as an input to an adversary network) and s-variant (i.e., z_n used as an input to a nuisance network) respectively. (b) Disentangled conditional rateless autoencoder (DA-cRAE) with soft split: a cRAE model with soft disentanglement, where the adversary and nuisance network inputs are determined through the stochastic bottleneck architecture with probabilities $p_a(d)$ and $p_n(d)=1-p_a(d)$ respectively for the dth latent node.

Unlike a conventional AE with a deterministic bottleneck architecture, the RAE employs a probabilistic bottleneck feature z whose dimensionality D is stochastically reduced through dropout. Particularly, RAE imposes a specific dropout rate distribution that varies across the D nodes of representation z. For example, as depicted in Fig. 1(b), the RAE encoder generates latent variables z of dimension D which are randomly dropped out at a probability of $p(d)$ for node $d \in \{1,2,\ldots,D\}$, resulting in an effective latent dimensionality of $\bar{D}=\sum_{d=1}^{D}(1-p(d))$. RAE is regarded as an ensemble method which jointly exploits all different AEs having a latent dimension of d from 1 to D. It is hence more insensitive to the choice of the dimensionality parameter.

In our method, we make use of the RAE concept to realize a good trade-off between task-related features and person-discriminative information by attaching new adversary and nuisance blocks to the representation z through different dropout strategies, with z fed into the decoder without dropout. A soft-disentangled feature extractor is first trained based on the rateless conception, and a task classifier is then learned for the final discriminative model utilizing the features extracted from the pre-trained (frozen) feature encoder.

C. Disentangled Adversarial Transfer Learning with RAE

1) Disentangled Feature Extractor: In [18] and [19], a disentangled feature extraction method was proposed to improve subject-transfer performance as shown in Fig. 2(a) (note that in [18] a task classifier instead of a decoder was attached to the encoder directly), where the features z are divided into two parts of z_a and z_n, which are intended to conceal subject-invariant and subject-specific information, respectively. Despite the gain of the disentangled method, determining the hard-split sizes of z_a and z_n is still challenging when the relationship between task- and subject- related features is unknown, which would also cause an additional hyperparameter that needs to be optimized. In this paper, we extend the method with soft disentanglement motivated by RAE as shown in Fig. 2(b), to mitigate the sensitivity of the splitting parameter, reduce the calculation and seamlessly adjust the smooth transition between task- and subject- relevant features.

For implementing the soft-disentangled adversarial transfer learning, encoder output z is forwarded into two additional units, the *adversary network* and *nuisance network*, with different dropout rate distributions. As illustrated in Fig. 2(b), the dropout rate distributions of representation z to the adversary network and nuisance network are designed as $p_a(d)$ and $p_n(d)=1-p_a(d)$, respectively. Complete latent representation z is further fed into the decoder $h(z,s;\eta)$ without any dropout. Through the stochastic disentangling, the representations z are re-organized into two sub-parts related to task and subject respectively: upper feature units with lower $p_a(d)$ (higher $p_n(d)$) to adversary network aim to conceal more subject information regarding s, while lower units with lower $p_n(d)$ (higher $p_a(d)$) to nuisance network are designed to include more subject-related features. By dissociating the nuisance variable from task-related feature in a more clear way, the model is extrapolated into a broader domain of subjects and tasks. For the input data from an unknown user, task-related features with lower $p_a(d)$ would be incorporated into the final prediction; simultaneously, the biological characteristics

which are similar to known subjects could also be projected to representations with lower $p_n(d)$ as a reference.

In order to filter out the variation elements caused by s from the adversary counterpart of z with lower $p_a(d)$ and simultaneously maintain more task-relevant information in it, the encoder is driven to minimize the adversary likelihood of $q_\phi(s|z,p_a)$; at the same time, to embed sufficient user-discriminative features within representations with lower $p_n(d)$, the encoder is also forced to maximize the nuisance likelihood of $q_\psi(s|z,p_n)$. The full representation z from encoder is fed into the decoder with zero dropout, which is conditioned on s as an additional input besides z, where the encoder and decoder are trained to optimize the reconstruction loss of $\hat{X}$ compared to the true input X. Therefore, the final objective function to train the proposed model structure can be written as follows:

$$\begin{aligned}\mathrm{Loss}_{\mathrm{RAE}}(X;\eta,\theta,\psi,\phi) = &-\mathbb{E}\big[\log p_\eta\big(\hat{X}|g(X;\theta),s\big)\big] \\ &- \lambda_N \mathbb{E}\left[\log q_\psi(s|z,p_n)\right] + \lambda_A \mathbb{E}\left[\log q_\phi(s|z,p_a)\right], \end{aligned} \quad (1)$$

where the first item is the loss of decoder $\hat{X} = h(z,s;\eta)$ reconstructing inputs from $z = g(X;\theta)$, and $\lambda_A \geq 0$ and $\lambda_N \geq 0$ respectively represent the regularization weights for adversary and nuisance units in order to achieve a flexible trade-off between identification and invariance performance. The model will reduce to a regular conditional AE (cAE) structure when $\lambda_A = \lambda_N = 0$, which involves no stochastic bottleneck or disentangling transfer learning block.

2) Adversarial Training Scheme: In addition to the training of encoder-decoder pair, at every optimization iteration, the weights of adversary and nuisance networks are learned towards maximizing the likelihoods $q_\phi(s|z,p_a)$ and $q_\psi(s|z,p_n)$ respectively to estimate the ID s among S subjects. The parameter updates and optimizations among the encoder-decoder pair, adversary network and nuisance network are performed alternatingly by stochastic gradient descent, where the two adversarial discriminators are separately trained to minimize their corresponding cross-entropy losses.

3) Discriminative Classifier: An independent status/task classifier is attached to the encoder with frozen network weights pre-trained by the proposed soft-disentangled adversarial method, and then optimized utilizing the input of latent feature z. The purpose of the classifier is to estimate the corresponding status/task class y among L categories given the physiological input X, where the feature $z = g(X;\theta)$ of X would be first extracted ahead to the task classifier. Parameterized by γ, the classifier optimization is further executed by minimizing the following cross-entropy loss:

$$\mathrm{Loss}_{\mathrm{C}}(z;\gamma) = \mathbb{E}\left[-\log p_\gamma(\hat{y}|z)\right], \quad (2)$$

where $\hat{y}$ is the estimate of subject status/task category y.

D. Discussion of Dropout Rate Distribution

Within the various dropout rate distributions for the representation z input to the adversary and nuisance networks (when $\lambda_A \geq 0$ and $\lambda_N \geq 0$), the proposed stochastic bottleneck architecture shown in Fig. 2(b) includes four cases: baseline AE, adversary unit only, hard split and soft split.

1) Baseline AE: When the dropout rates $p_a(d) = p_n(d) = 1$ and $\lambda_A = \lambda_N = 0$ for all feature nodes $d \in z$, feature z is not connected to any of the adversary and nuisance blocks. Therefore, the model reduces to a baseline AE architecture with a regular encoder-decoder pair for feature extraction as presented in [25] and [26], whose decoder is $h(z;\eta)$ without adversarial disentangling units. We also denote cAE as a conditional AE feature extractor with decoder $h(z,s;\eta)$ conditioned on s as described in [27].

2) Adversary Unit Only: While the dropout rate $p_a(d) = 0$ with $\lambda_A > 0$ and simultaneously $p_n(d) = 1$ with $\lambda_N = 0$ for every node d, the entire feature group z is directly linked to the adversary network without both feature split and connection to a nuisance network. The same operation can be applied to the nuisance network when $p_a(d) = 1, \lambda_A = 0$ and $p_n(d) = 0, \lambda_N > 0$. Here we utilize A-cAE and D-cAE to denote the cAE models with only the adversary/nuisance network attached. Note that the A-cAE resembles to the traditional adversarial learning methods presented in [20–24] where only one adversarial unit is adopted.

3) Hard Split: For the particular case when the dropout rate $p_a(d) = 1 - p_n(d)$ is either 0 or 1 for each feature node d, i.e., when the feature output of node d is either input to the adversary network only or the nuisance network only along with decoder, the representation z is hard split into two sub-parts z_a and z_n, corresponding respectively to the adversary and nuisance blocks, as shown in Fig. 2(a). The sub-part feature z_a with $p_a(d) = 0$ and $p_n(d) = 1$ for $d \in z_a$ aims at preserving task-related feature information, while subject-related feature would be embedded in representation z_n with $p_a(d) = 1$ and $p_n(d) = 0$ for $d \in z_n$. In this case, it reduces to a regular disentangled adversarial cAE structure (DA-cAE) with adversary and nuisance networks attached but no rateless property. This structure corresponds to the method in [18] and [19], which is a special case of the following soft-split method with different dropout rate distributions.

4) Soft Split: For the more generic case of soft-split representation z, dropout rates to adversary and nuisance blocks are arbitrary, provided that they satisfy $p_a(d) = 1 - p_n(d) \in [0,1]$ for each feature node $d \in \{1,2,\ldots,D\}$. Therefore, the bottleneck architecture z is soft split into adversary and nuisance counterparts stochastically according to the distribution $p_a(d)$ and $p_n(d) = 1 - p_a(d)$, respectively, as depicted in Fig. 2(b). This conditional RAE with soft-disentangled adversarial structure (DA-cRAE) can partly resolve the issue of hard split which requires pre-determined dimensionality for two disentangled latent vectors, whereas the proposed method can automatically consider different ratio of hard splits in a non-deterministic ensemble manner.

E. Model Implementations

Motivated by recent impressive results for biosignal processing using deep learning tools [29], [30], we mainly make use of neural networks to build the feature extractor in the proposed model. However, we note that other learning frameworks without neural networks could also be applied to the proposed method of soft-disentangled adversarial transfer learning.

TABLE I: Network structures, where FC($d_\mathrm{i}, d_\mathrm{o}$) is linear fully connected layer of dimensions d_i and d_o for input and output, and ReLU denotes rectified linear unit.

Encoder Network	FC(C, D) $\rightarrow$ ReLU $\rightarrow$ FC(D, D)
Decoder Network	FC(D, D) $\rightarrow$ ReLU $\rightarrow$ FC(D, C)
Adversary Network	FC(D, S)
Nuisance Network	FC(D, S)

1) Model Architecture: The utilized model structure for experiment evaluations is presented in Table I, where representation z has a dimensionality of D. The adversary and nuisance networks have a same input dimension D from the latent representation and output dimension S for the classification of subject IDs. We note that we did not observe significant improvements by deepening the network or altering the number of units for our physiological biosignal dataset under test. To assess the robustness of the proposed soft-disentangled adversarial feature encoder, we implemented various classifiers for evluating the final task classification, including MLP, nearest neighbors, decision tree, linear discriminant analysis (LDA), and logistic regression classifiers with L output dimensions for task classification.

2) Rateless Parameters: Representation z with dimension $D = 15$ is fed into adversary network and nuisance network respectively with dropout rates $p_a(d)$ and $p_n(d)$. For the soft-split case in Section II-D.4, we take $p_a(d) = ((d-1)/(D-1))^\alpha$ and $p_n(d) = 1 - p_a(d)$ for $d \in \{1, 2, \ldots, D\}$, where parameter α can adjust the ascent speed of dropout rate $p_a(d)$ along d, and we take $\alpha = 3$ in the experimental assessments, since we need relatively more weight from the task-related features than the nuisance-variable-relevant counterpart during the task classification. In the implementation for hard split of Section II-D.3, we fix the ratio of dimensions between z_a and z_n to $2:1$.

3) Comparison Model Definitions: We denote AE as a baseline architecture of a regular encoder-decoder pair for feature extraction as presented in [25] and [26], whose decoder is $h(z;\eta)$ without adversarial disentangling units, and cAE as a conditional AE feature extractor with decoder $h(z, s;\eta)$ conditioned on s as described in [27]. A-cAE and A-cRAE denote the cAE models with the aforementioned hard-split and soft-split bottleneck features respectively attached to the adversary network only. D-cAE and D-cRAE represent cAE with hard-split and soft-split bottleneck variables respectively linked to the nuisance network only. DA-cAE and DA-cRAE specify hard-split and soft-split representations connected to both adversary and nuisance networks respectively with decoder conditioned on s. Note that the A-cAE resembles to the traditional adversarial learning methods presented in [20–22], [24] where only one adversarial unit is adopted.

III. Experimental Study

A. Dataset

The proposed methodology was evaluated on a physiological biosignal dataset for assessing human stress status [9], which is available online[1]. It includes physiological biosignals of various modalities, in order to estimate $L = 4$ discrete stress levels (physical stress, cognitive stress, emotional stress, and relaxation) based on data collected from $S = 20$ subjects. The biosignals were generated from non-invasive biosensors worn on the wrist, containing heart rate, temperature, electrodermal activity, three-dimensional acceleration, and arterial oxygen level, therefore resulting in $C = 7$ signal channels totally. We further downsampled the signals to 1 Hz in order to align all data channels. For each stress status, a 5-minute long task was assigned to the subjects. In total, 7 trials were executed by every subject, among which 4 trials were the status of relaxation. To address the data imbalance of trials with different categories, we only utilized the first trial of relaxation status, leading to four trials for the four stress status levels respectively and 24,000 data samples in total.

B. Experiment Implementation and Evaluation

We implemented all models in Python3 using the Chainer framework, which is an open-source deep learning framework. The regularization weights λ_A and λ_N were chosen for the disentangled adversarial model by parameter sweep and cross-subject validation. We trained the feature extractor with different parameter combinations, and preferred the parameters producing lower accuracy of the adversary discriminator and higher accuracy of the nuisance discriminator, premised on obtaining higher cross-validation accuracy for the discriminative task classifier trained by extracted features. We evaluated the model with transfer analysis of cross-subjects through a leave-one-subject-out method [6], where the cross-subject test data came from the left-out subject.

For each held-out subject, we first learn the encoder-decoder pair built with neural network from scratch, and then optimize the task classifier using the trained neural-network encoder. During the training of the encoder, 90% and 10% of the data from the remaining 19 subjects were randomly split as the training and validation sets for optimizing the neural network only, where the validation set was used for early stopping the training to prevent overfitting, i.e. the network training would be terminated as soon as the performance on the validation dataset decreases as compared to the performance on the validation dataset at the prior training epoch. Then an independent task classifier attached to the pre-trained encoder with frozen network weight would be optimized utilizing the latent features from the training data of the remaining subjects. Finally, the classification accuracy would be evaluated on the left-out subject. The described training and testing process was repeated for each held-out subject (20 times in total), and the overall accuracy was averaged over all cross-subjects accuracies.

The λ_A and λ_N values were selected and optimized based on higher averaged cross-subjects testing accuracies over all held-out subjects, since a global accuracy could better reflect the generalization and robustness of the model on different individuals. To reduce the size of parameter combinations, we first swept over λ_N with $\lambda_A = 0$; then

[1] https://physionet.org/content/noneeg/1.0.0/

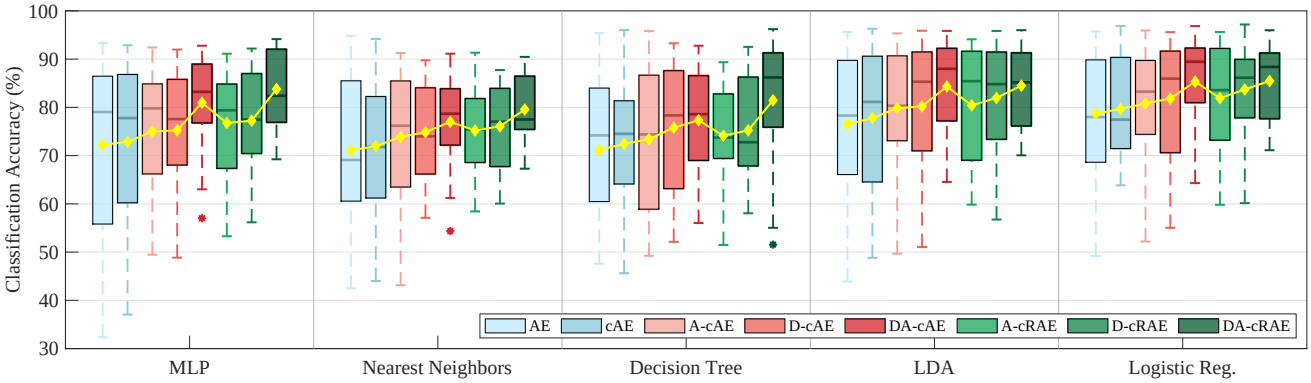


Fig. 3: Averaged cross-subjects accuracies of different classifiers with eight feature learning frameworks: (1) AE: baseline of regular AE with decoder $h(z;\eta)$, (2) cAE: AE with s-conditional decoder $h(z,s;\eta)$, (3) A-cAE: hard-split bottleneck cAE with adversary network, (4) D-cAE: hard-split bottleneck cAE with nuisance network, (5) DA-cAE: hard-split bottleneck cAE with both adversary and nuisance networks, (6) A-cRAE: soft-split bottleneck cAE with adversary network, (7) D-cRAE: soft-split bottleneck cAE with nuisance network, (8) DA-cRAE: soft-split bottleneck cAE with both adversary and nuisance networks. For each box, the central line marks the median, upper and lower bounds represent first and third quartiles, and dashed lines denote extreme values; the diamond-shape marker specifies the average.

TABLE II: Optimized parameter selections with averaged cross-subjects accuracies.

	MLP			Nearest Neighbors			Decision Tree			LDA			Logistic Regression		
	λ_A	λ_N	avg acc	λ_A	λ_N	avg acc	λ_A	λ_N	avg acc	λ_A	λ_N	avg acc	λ_A	λ_N	avg acc
AE [25], [26]	0	0	72.2%	0	0	71.1%	0	0	71.2%	0	0	76.5%	0	0	78.7%
cAE [27]	0	0	72.9%	0	0	72.2%	0	0	72.4%	0	0	77.8%	0	0	79.7%
A-cAE [20–22], [24]	0.005	0	75.0%	0.1	0	73.9%	0.1	0	73.4%	0.05	0	79.8%	0.05	0	80.8%
D-cAE	0	0.005	75.2%	0	0.01	74.9%	0	0.01	75.8%	0	0.2	80.2%	0	0.2	81.8%
DA-cAE [18], [19]	0.01	0.005	**81.0%**	0.1	0.01	**77.0%**	0.2	0.01	**77.3%**	0.2	0.2	**84.3%**	0.2	0.2	**85.3%**
A-cRAE	0.02	0	76.8%	0.05	0	75.2%	0.05	0	74.1%	0.1	0	80.4%	0.02	0	81.9%
D-cRAE	0	0.05	77.2%	0	0.05	76.1%	0	0.1	75.2%	0	0.05	82.0%	0	0.05	83.7%
DA-cRAE	0.5	0.05	**83.8%**	0.5	0.05	**79.6%**	0.01	0.1	**81.5%**	0.5	0.05	**84.5%**	0.5	0.05	**85.5%**

λ_N was fixed at its optimized value from the previous step to optimize λ_A value. The adopted ranges of λ_A and λ_N are $\lambda_A \in \{0, 0.01, 0.05, 0.1, 0.2, 0.5\}$ and $\lambda_N \in \{0, 0.005, 0.01, 0.05, 0.2, 0.5\}$. Note that the selected parameter values can be even optimized more within larger scopes by cross-validating the same model learning process.

C. Results and Discussions

1) Comparative Experiments: Averaged accuracies of transfer analysis across 20 held-out subjects based on different feature encoders and classifiers are presented in Fig. 3, where AE, cAE, A-cAE, D-cAE, DA-cAE, A-cRAE, D-cRAE, and DA-cRAE as defined in Section II-E.3 were trained and compared. Corresponding parameter settings for each case in Fig. 3 are displayed in Table II, which were selected and optimized through the aforementioned parameter optimization procedure. The model architecture is as shown in Table I, where feature dimension is $D = 15$.

As shown in Fig. 3 and Table II, first we observe that simply feeding the decoder an extra conditional input s could yield slightly better classification performance when comparing cAE with AE. Furthermore, we notice accuracy improvements from A-cAE and D-cAE to cAE, demonstrating that more cross-subject features observed in the hard-split representation z_a lead to better identification of y. In addition, DA-cAE realizes further accuracy improvements with both adversary and nuisance networks compared to individual regularization approaches A-cAE and D-cAE. Under the disentangled adversarial transfer learning framework, our feature extractor results in lower variation of performances across all task classifiers and all subjects universally. More importantly, the soft-split

TABLE III: Parameter optimization of MLP classifier. Accuracies for the adversary, nuisance and classifier are presented.

	λ_A	λ_N	**MLP Classifier**	**Adversary Network**	**Nuisance Network**
AE	0	0	72.2%	7.8%	5.6%
cAE	0	0	72.9%	8.5%	5.8%
D-cRAE	0	0.005	74.8%	7.7%	8.5%
	0	0.01	73.5%	12.5%	15.2%
	0	**0.05**	**77.2%**	**10.7%**	**19.7%**
	0	0.2	75.6%	13.6%	16.5%
	0	0.5	74.1%	12.6%	35.5%
DA-cRAE	0.01	0.05	78.3%	9.4%	13.6%
	0.05	0.05	77.3%	6.7%	14.6%
	0.1	0.05	77.9%	5.9%	13.3%
	0.2	0.05	81.5%	5.5%	12.7%
	0.5	**0.05**	**83.8%**	**4.9%**	**13.9%**

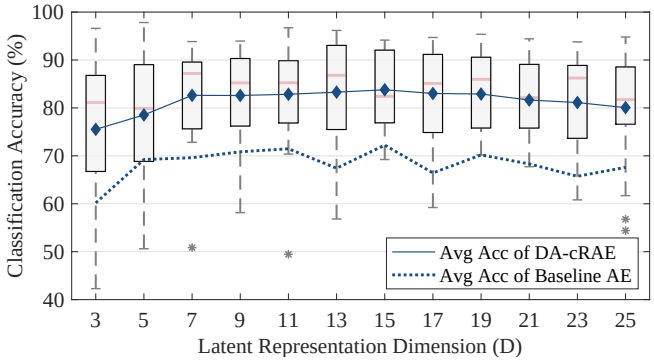


Fig. 4: MLP classification accuracies of DA-cRAE model ($\lambda_A = 0.5$, $\lambda_N = 0.05$) for 20 held-out subjects with different dimension D of representation z, compared with baseline AE.

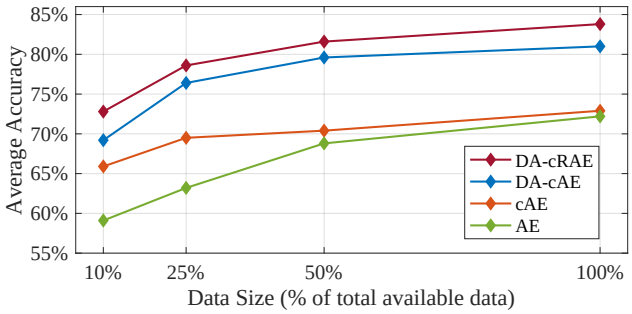


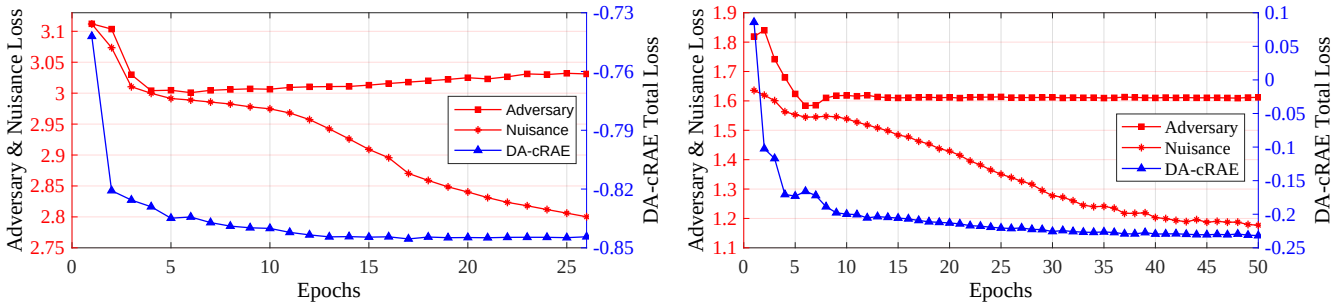
Fig. 5: MLP classification accuracies of optimized parameter choices in Table II with different training dataset sizes.

RAE structures of A-cRAE, D-cRAE and DA-cRAE bring even more accuracy gain compared to the hard-split cases of A-cAE, D-cAE and DA-cAE. For the hard-split case, determining the split ratio of dimensions between subject-related and task-specified features is difficult since the representation nature is still unkown. However, the rateless property enables the encoder-decoder pair to seamlessly adjust dimensionalities of subject-related and task-specified features, and employs a smooth transition between the two stochastic counterparts by a probabilistic bottleneck representation, even though the underlying nature of the bottleneck is still vague.

In general, the disentangled adversarial models of DA-cRAE with both adversary and nuisance networks attached to conditional decoder lead to significant improvements in average accuracy up to 11.6% (e.g., the MLP classifier in Table II) with respect to the non-adversarial baseline AE. Note that the statistical significance of our DA-cRAE superiority over the baseline AE (and DA-cAE with the exception of LDA and Logistic Regression) is confirmed through the independent-sample t-test with a significance level of $p < 10^{-13}$. Furthermore, as observed in Fig. 3, the cross-validation accuracies of the worst cases are also significantly improved, indicating that the proposed transfer learning architecture presents higher stability to a wider range of unknown individuals through reorganizing the subject- and task-relevant representations from the end of feature extractor. It is worth noting that, for different classifiers, the selection of the optimized λ_A and λ_N combination could be different due to the distinct characteristics of classifiers when classifying input data, for example the Logistic Regression is a linear method whereas the Decision Tree is based on a tree structure. Thus comprehensive experiments for a wide range of parameters are necessary under different application scenarios.

2) Impact of Disentangled Adversarial Parameters: We take the MLP classifier as an example to particularly illustrate the impact of disentangled adversarial RAE. As presented in Table III, the baseline models of AE and cAE were first assessed with $\lambda_A = \lambda_N = 0$ while training the MLP discriminative classifier. Then the D-cRAE was evaluated with $\lambda_N \in \{0, 0.005, 0.01, 0.05, 0.2, 0.5\}$ and $\lambda_A = 0$. Finally, we froze $\lambda_N = 0.05$ to observe the representation learning capability of the complete soft-disentangled adversarial transfer learning model DA-cRAE with different choices of $\lambda_A \in \{0, 0.01, 0.05, 0.1, 0.2, 0.5\}$. For each parameter selection, the average accuracy of the MLP task classifier for identifying 4 stress levels is shown in Table III, along with the discriminator accuracies of the adversary and nuisance blocks for decoding 20-class ID. With an increasing accuracy of MLP task classifier, stress levels are better discriminated; with a growing accuracy of nuisance network, more person-discriminative features are preserved in the nuisance counterpart; and with a decreasing accuracy of adversary network, more task-specific information are inherent in the adversary counterpart. We observe that the nuisance network produces higher accuracy with increasing λ_N, where $\lambda_N = 0.05$ particularly results in the better performance on task classification. Furthermore, with fixed $\lambda_N = 0.05$, growing λ_A leads to lower accuracy of adversary network, and thus imposes less extraction of subject features but more task-related information on the adversary counterpart.

3) Impact of Feature Dimension: Other than the adversarial parameters λ_A and λ_N, we further inspect the impact of different feature dimensions D on the performance of the proposed DA-cRAE model. We trained MLP classifiers with the DA-cRAE feature extractor and its optimized parameters as given in Table II ($\lambda_A = 0.5$ and $\lambda_N = 0.05$), using various feature dimensions $D \in \{3, 5, \cdots, 25\}$. Corresponding cross-validation accuracies for 20 held-out subjects are shown as a function of D in Fig. 4, where the average accuracy for each D is also marked. The same assessments on D were also applied to baseline AE feature extractor, and we present its curve of average accuracies in Fig. 4 as a reference to compare with DA-cRAE. It is verified that the proposed DA-cRAE consistently outperforms the baseline AE and $D = 15$ latent dimensionality was sufficient for the problem. We observe that after a specific value of dimension D, the performance of DA-cRAE remains relatively stable with varying D value compared to AE. On one hand, when the feature dimension is large enough to carry necessary information for the classification task, higher D value might not be able to bring more benefits when extracting features; on the other hand, the rateless property of DA-cRAE resolves the entanglement between task-related and subject-discriminative information and exploits the latent features in



(a) DA-cRAE convergence curves with 100% data size.

(b) DA-cRAE convergence curves with 25% data size.

Fig. 6: Convergence of DA-cRAE ($\lambda_A = 0.5$ and $\lambda_N = 0.05$) with different training data sizes.

a more efficient manner, thus leading to a stronger robustness on the variance of latent representation dimensionality.

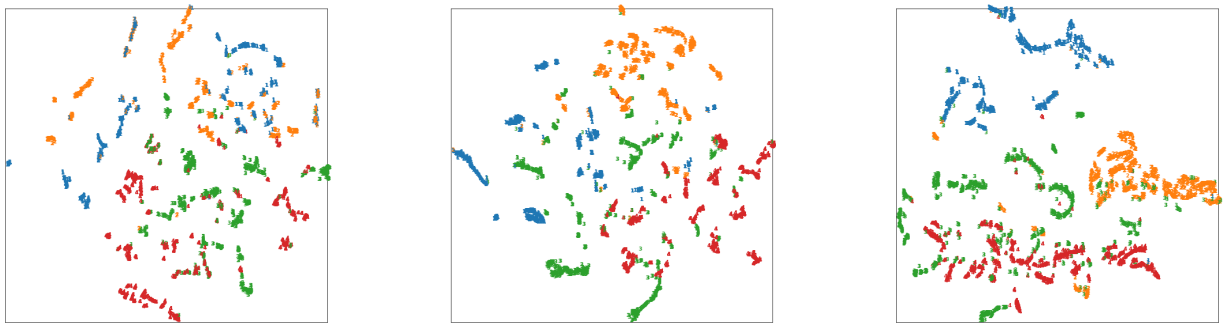
4) Impact of Data Size: In order to evaluate the robustness of our transfer learning method on data with smaller sizes, we investigated the performance of the proposed model when we reduced the available training data size from 100% to 50%, 25%, or 10%. Corresponding classification accuracies as a function of training data size are shown in Fig. 5. Here we consider the MLP classifier as the same example of Table II, to make comparisons among DA-cRAE ($\lambda_A = 0.5$ and $\lambda_N = 0.05$), DA-cAE ($\lambda_A = 0.01$ and $\lambda_N = 0.005$), cAE and AE ($\lambda_A = \lambda_N = 0$). From Fig. 5 we observe that DA-cRAE still performs best regardless of the amount of available training data. Even with 10% data only, there is no significant drawback of DA-cRAE and DA-cAE compared to non-adversarial methods, showing the transfer learning ability of our method to the size deficiency of physiological data. Note that with more available training data, even better performance is expected to be implemented by the model.

5) Convergence Analysis: In addition, training convergence curves for a specific DA-cRAE ($\lambda_A = 0.5$ and $\lambda_N = 0.05$) case with different training data sizes are presented in Fig. 6. When using the full 100% set of available training data, i.e., in Fig. 6(a), the total training loss value of DA-cRAE converges within 15 epochs, while the nuisance loss decreases steadily with more training iterations and the loss value of the adversary unit keeps steady due to its antagonistic relationship with DA-cRAE, where the adversary unit continues to conceal subject-specific representations without undermining the discriminative performance of the entire network. With less data, as illustrated in Fig. 6(b), convergences are achieved after more training epochs, while the convergences of the DA-cRAE loss, adversary loss and nuisance loss are observed in a similar pattern with the full 100% data case, indicating the capability of the proposed model to learn universal features from data with even smaller sizes. Overall, we observe that with both adversary and nuisance networks attached to the encoder, the classifier improves the accuracy substantially and shows more stable performance across different left-out subjects.

6) Feature Visualization and Analysis: In order to inspect the features extracted by the DA-cRAE encoder, we use the t-distributed stochastic neighbor embedding (t-SNE) [31] for visualizing the high-dimensional latent features in a two-dimensional map. The t-SNE is a popular method to map a high-dimensional data into a low-dimensional data in favor of minimizing the Kullback–Leibler divergence, so that the input data could be visualized in a more intuitive space for clustering. In Fig. 7, the t-SNE map with respect to a four-class task category y is shown, for the raw multi-channel physiological data, the features extracted by baseline AE model, and the features extracted by DA-cRAE model. From Fig. 7, we can observe that the features generated by our DA-cRAE model can be distinguished by different tasks relatively better than the raw data and baseline AE features, indicating an improved discriminating ability of the proposed model.

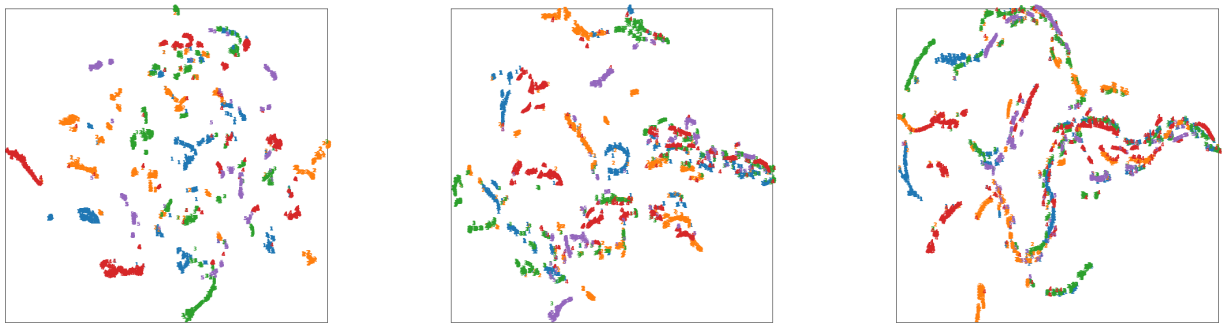
We next demonstrate the soft disentanglement of the adversarial framework for nuisance-robust learning. The DA-cRAE latent features (dimension $D = 15$) were soft-disentangled into subject-invariant and subject-dependent counterparts during the adversarial training, where the two counterparts are connected seamlessly. Based on the dropout rate distribution used in Section II-E.2, we consider the first ten latent features ($d = 1, 2, \ldots, 10$) as subject-invariant counterpart since it has relatively lower dropout rate to the adversary network, while regard the rest five features ($d = 11, 12, \ldots, 15$) as subject-relevant counterpart which is connected to the nuisance network with lower dropout rates. In Fig. 8, the t-SNE maps of those two counterparts of DA-cRAE features regarding to the nuisance variable s (subject IDs) are presented, along with the baseline AE features.

As presented in Fig. 8(a), the features extracted from the baseline AE still retain the same spatial distribution as shown in Fig. 7(b), whereas in Fig. 8 the samples are annotated by subject IDs instead of task labels. In Fig. 8(b) and (c), the 10-dimensional task-related counterpart and 5-dimensional subject-relevant counterpart of the entire 15-dimensional DA-cRAE features are respectively displayed and marked by different subject IDs. From Fig. 8(b) we observe that the sample layout (despite the distinct annotation colors) of the task-related feature map is highly similar to the entire DA-cRAE representations as shown in Fig. 7(c), illustrating the neat selection and disentanglement of the nuisance-invariant features which also turn to be more contributing to the task classification. Furthermore, compared to Fig. 8(a) and (b), the map of the subject-related features in Fig. 8(c) shows a distinct characteristics which is more separable in regards to



(a) Raw data (b) Features from baseline AE (c) Features from DA-cRAE

Fig. 7: The t-SNE visualization with respect to the four-class task category y: (a) raw multi-channel physiological data; (b) features extracted by baseline AE model; and (c) features extracted by DA-cRAE model ($\lambda_A = 0.5$, $\lambda_N = 0.05$).



(a) Features from baseline AE (b) Task-related features of DA-cRAE (c) Subject-related features of DA-cRAE

Fig. 8: The t-SNE visualization of the baseline AE features, and the task-/subject-related counterparts of 15-dimensional DA-cRAE features (with $\lambda_A = 0.5$, $\lambda_N = 0.05$) with respect to the nuisance variable s (subject IDs): (a) the entire 15-dimensional features extracted by baseline AE model; (b) the 10-dimensional task-related counterpart of DA-cRAE features, which has lower dropout rate to adversary network; and (c) the 5-dimensional subject-related counterpart of DA-cRAE features, which has lower dropout rate to nuisance network.

subject IDs and thus can provide more information about the users. The subject-specific representations will eventually help the task-dependent counterpart to stand out from all features through their adversarial relationship, while maintaining the robustness for transfer learning in a seamless manner instead of distracting the system.

IV. Conclusion

A transfer learning framework was proposed based on a soft-disentangled adversarial model utilizing the concept of RAE to extract universal and nuisance-robust physiological features. In order to implement the rateless property and manipulate the trade-off between subject-specific features and task-relevant information, additional blocks of adversary and nuisance networks were complementarily attached and jointly trained with different dropout strategies, and therefore the transfer learning framework is capable of handling a wider range of tasks and users. Cross-subject transfer evaluations were performed with a physiological biosignal dataset for monitoring human stress levels. Significant benefits of the proposed framework were shown by improved worst-case accuracy and average classification accuracy, demonstrating the robustness to unknown users. The adaptability of the feature extractor over several task-discriminative linear and non-linear classifiers was also shown, and the transfer-learning ability of our method to data size deficiency was analysed. Note that our methodology is applicable to various different systems requiring nuisance-robust analysis beyond HCI.

References

[1] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *IEEE 7th International Colloquium on Signal Processing and its Applications*. IEEE, 2011, pp. 410–415.

[2] S. H. Jambukia, V. K. Dabhi, and H. B. Prajapati, "Classification of ecg signals using machine learning techniques: A survey," in *2015 International Conference on Advances in Computer Engineering and Applications*. IEEE, 2015, pp. 714–721.

[3] M. Han, S. Y. Günay, G. Schirner, T. Padır, and D. Erdoğmuş, "Hands: a multimodal dataset for modeling toward human grasp intent inference in prosthetic hands," *Intelligent Service Robotics*, vol. 13, no. 1, pp. 179–185, 2020.

[4] M. Han, S. Y. Günay, İ. Yildiz, P. Bonato, C. D. Onal, T. Padir, G. Schirner, and D. Erdoğmuş, "From hand-perspective visual information to grasp type probabilities: deep learning via ranking labels," in *Proceedings of the 12th ACM international conference on pervasive technologies related to assistive environments*, 2019, pp. 256–263.

[5] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 186–197, 2009.

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

[7] Z. Jiao, X. Gao, Y. Wang, J. Li, and H. Xu, "Deep convolutional neural networks for mental load classification based on EEG data," *Pattern Recognition*, vol. 76, pp. 582–595, 2018.

[8] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[9] J. Birjandtalab, D. Cogan, M. B. Pouyan, and M. Nourani, "A non-EEG biosignals dataset for assessment and visualization of neurological status," in *IEEE International Workshop on Signal Processing Systems*, 2016, pp. 110–114.

[10] A. M. Amiri, M. Abtahi, A. Rabasco, M. Armey, and K. Mankodiya, "Emotional reactivity monitoring using electrodermal activity analysis in individuals with suicidal behaviors," in *10th International Symposium on Medical Information and Communication Technology*, 2016, pp. 1–5.

[11] D. Cogan, M. B. Pouyan, M. Nourani, and J. Harvey, "A wrist-worn biosensor system for assessment of neurological status," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 5748–5751.

[12] D. Giakoumis, D. Tzovaras, and G. Hassapis, "Subject-dependent biosignal features for increased accuracy in psychological stress detection," *International Journal of Human-Computer Studies*, vol. 71, no. 4, pp. 425–439, 2013.

[13] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Transactions on Affective Computing*, 2019.

[14] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, and C. Grozea, "Subject-independent mental state classification in single trials," *Neural Networks*, vol. 22, no. 9, pp. 1305–1312, 2009.

[15] H. Morioka, A. Kanemura, J.-I. Hirayama, M. Shikauchi, T. Ogawa, S. Ikeda, M. Kawanabe, and S. Ishii, "Learning a common dictionary for subject-transfer decoding with resting calibration," *NeuroImage*, vol. 111, pp. 167–178, 2015.

[16] Z. Yin, M. Zhao, W. Zhang, Y. Wang, Y. Wang, and J. Zhang, "Physiological-signal-based mental workload estimation via transfer dynamical autoencoders in a deep learning framework," *Neurocomputing*, vol. 347, pp. 212–229, 2019.

[17] L.-L. Chen, A. Zhang, and X.-G. Lou, "Cross-subject driver status detection from physiological signals based on hybrid feature selection and transfer learning," *Expert Systems with Applications*, vol. 137, pp. 266–280, 2019.

[18] M. Han, O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Disentangled adversarial transfer learning for physiological biosignals," in *42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2020.

[19] ——, "Disentangled adversarial autoencoder for subject-invariant physiological feature extraction," in *IEEE Signal Processing Letters*, 2020.

[20] H. Edwards and A. Storkey, "Censoring representations with an adversary," *arXiv preprint arXiv:1511.05897*, 2015.

[21] F. Wu, X.-Y. Jing, Z. Wu, Y. Ji, X. Dong, X. Luo, Q. Huang, and R. Wang, "Modality-specific and shared generative adversarial network for cross-modal retrieval," *Pattern Recognition*, p. 107335, 2020.

[22] Y. Sun, X.-Y. Jing, F. Wu, J. Li, D. Xing, H. Chen, and Y. Sun, "Adversarial learning for cross-project semi-supervised defect prediction," *IEEE Access*, vol. 8, pp. 32 674–32 687, 2020.

[23] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Learning invariant representations from EEG via adversarial inference," *IEEE Access*, vol. 8, pp. 27 074–27 085, 2020.

[24] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato, "Fader networks: Manipulating images by sliding attributes," in *Advances in Neural Information Processing Systems*, 2017, pp. 5967–5976.

[25] M. Tschannen, O. Bachem, and M. Lucic, "Recent advances in autoencoder-based representation learning," *arXiv preprint arXiv:1812.05069*, 2018.

[26] T. Wen and Z. Zhang, "Deep convolution neural network and autoencoders-based unsupervised feature learning of eeg signals," *IEEE Access*, vol. 6, pp. 25 399–25 410, 2018.

[27] Y. Yang, K. Zheng, C. Wu, and Y. Yang, "Improving the classification effectiveness of intrusion detection by using improved conditional variational autoencoder and deep neural network," *Sensors*, vol. 19, no. 11, p. 2528, 2019.

[28] T. Koike-Akino and Y. Wang, "Stochastic bottleneck: Rateless auto-encoder for flexible dimensionality reduction," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020.

[29] M. Atzori, M. Cognolato, and H. Müller, "Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands," *Frontiers in Neurorobotics*, vol. 10, p. 9, 2016.

[30] O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya, "Deep learning for healthcare applications based on physiological signals: A review," *Computer Methods and Programs in Biomedicine*, vol. 161, pp. 1–13, 2018.

[31] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.