Model Adaptation and Personalization for Physiological Stress Detection

Aaqib Saeed[1], Tanir Ozcelebi[2], Johan Lukkien[3], Jan B.F. van Erp[4], Stojan Trajanovski[5]
{a.saeed, t.ozcelebi, j.j.lukkien}@tue.nl, jan.vanerp@utwente.nl, stojan.trajanovski@philips.com
[1,2,3]*Department of Mathematics and Computer Science*, *Eindhoven University of Technology*, Eindhoven, The Netherlands
[4]*Human Media Interaction, Computer Science*, *University of Twente*, Enschede, The Netherlands
[5]*Philips Research*, Eindhoven, The Netherlands

***Abstract*—Stress and accompanying physiological responses can occur when everyday emotional, mental and physical challenges exceed one's ability to cope. A long-term exposure to stressful situations can have negative health consequences, such as increased risk of cardiovascular diseases and immune system disorder. It is also shown to adversely affect productivity, well-being, and self-confidence, which can lead to social and economic inequality. Hence, a timely stress recognition can contribute to better strategies for its management and prevention in the future. Stress can be detected from multimodal physiological signals (e.g. skin conductance and heart rate) using well-trained models. However, these models need to be adapted to a new target domain and personalized for each test subject. In this paper, we propose a deep reconstruction classification network and multi-task learning (MTL) for domain adaption and personalization of stress recognition models. The domain adaption is achieved via a hybrid model consisting of temporal convolutional and recurrent layers that perform shared feature extraction through supervised source label predictions and unsupervised target data reconstruction. Furthermore, MTL based neural network approach with hard parameter sharing of mutual representation and task-specific layers is utilized to acquire personalized models. The proposed methods are tested on multimodal physiological time-series data collected during driving tasks, in both real-world and driving simulator settings.**

***Index Terms*—physiological stress, domain adaption, personalization, multi-task learning, deep learning, temporal convolutional neural networks**

I. Introduction

In today's society, we experience numerous stressful situations such as dealing with annual job evaluation, business failure, or illness. Stress is described as a psychophysiological response to mental, emotional and physical challenges encountered in daily life [1]. Even though the human body is capable of dealing with short-lived day-to-day stressors, the long-term exposure to unremitting stress can have destructive consequences for well-being, productivity, behavior, and self-confidence [2; 3]. Stress can also adversely affect health with implications for progression, recovery, and treatment of nearly every known disease through physiological, behavioral and cognitive changes [1]. It increases the risk of diabetes, metabolic disorders, cardiovascular diseases and (psycho) somatic complaints [4; 5]. Due to these health and performance issues, stress management becomes important. A timely detection of stress can be extremely powerful as it can empower users to take corrective and preventive measures in an informed manner [6].

The autonomic nervous system (ANS) consists of two branches, namely, sympathetic and parasympathetic nervous system which are both influenced by (amongst others) physiological stress and emotional arousal. The activity of the sympathetic part results in an increase in heart rate, blood pressure, respiration, and blood flow to the muscles. An activity of the parasympathetic division results in an increase in blood flow to the organs and the skin, a decrease in heart rate and respiration, and an increase in heart rate variability. The ANS responds to stress by stimulating specified target organs via efferent neuron tracts, initiated in the locus coeruleus of the brain stem [7] resulting in a release of noradrenaline and norepinephrine. The immediate effect thereof is an increase in sympathetic and a decrease in parasympathetic activity, resulting in a measurable change in physiological parameters, such as an increased heart rate (HR) and skin conductance (SC) level.

Assessing stress levels has a wide area of applications, from increasing resilience of military personnel to enhancing athletes' performance and improving workforce productivity. Several techniques have been proposed in the past to detect stress in pilots [8], car and truck drivers [9; 10; 11], computer users [12], call center employees [6] and in surgeons [13]. In addition to audio-visual modalities, most approaches use numerous physiological signals, such as respiration rate, electrocardiography (ECG), blood pressure, and electromyography (EMG). The collection of these data in natural conditions is very difficult and usually not consumer friendly enough for practical applications. In contrast, SC and HR can be reliably acquired in a non-invasive and non-obtrusive way from wearable sensors placed on the wrist. Currently, the key challenge is the reliable and personalized classification of stress-states based on these easy to obtain SC and HR signals. In the present work, we focus on personalization and unsupervised model adaption to improve stress assessment both inside and outside controlled lab environments (domains) using HR and SC signals.

The development of wearable sensors for electrodermal activity and heart rate monitoring has boosted the interest in using these for stress assessment over the last decade. Several recent works have shown their successful application

with machine learning algorithms to detect stress in different (mostly controlled) conditions [14; 15; 16]; see [17] for a detailed survey. The aspect that is evident from the overview of earlier work is that current methods do not address issues of end-to-end representation learning, covariate shift, personalization, and domain adaption. The traditional supervised learning algorithms are not robust to dataset bias [18] and may perform poorly when the data distribution of training instances (of a source domain) differs from test instances (of a target domain). For example, a model trained on a data collected in a simulated (constrained) environment may not be able to perform well in a real-world (unconstrained) setting. Hence, these methods require a collection of ground-truth data (in real-world) for model retraining and are unable to leverage unlabeled data directly to perform cross-domain stress classification. Similarly, physiological signals tend to vary in people and are influenced by age, gender, diet or sleep [19]. Due to this fact, stress responses can differ from person to person. The global (or one-fits-all) models, often do not generalize well to unseen test subjects and hence need extensive fine-tuning.

To address the aforementioned issues, we propose an end-to-end representation learning framework based on a deep reconstruction classification network [20] (DRCN) and multi-task learning (MTL). We focus on personalization and domain adaption together as DRCN can be seen as an extension of MTL. The objective of DRCN is to improve predictive performance on the target domain through joint training on labeled and unlabeled data points. It performs shared feature extraction through supervised source label predictions and unsupervised target data reconstruction. Specifically, the reconstruction phase enables the network to adapt the label prediction function for the target domain, which is similar to learning an auxiliary task in MTL setting to improve the performance on the actual task [21]. Likewise, model personalization can be achieved with MTL, if subjects are treated as tasks [22]. In this case, the multi-task neural network has hard (or soft) parameter sharing of mutual representations along with distinct layers for each subject (or task) to account for bodily interpersonal differences.

We demonstrate the versatility of the proposed methods via three datasets from a representative application area i.e. stress detection in a driving context. Our approach makes no assumption about the sensor types, sampling frequency, and structure of the physiological time series. It is important to note that these methods are flexible, they can be applied to a variety of neural network architectures and can be used for a variety of different time series classification tasks with minimal changes. Additionally, DRCN and MTL models learned in an end-to-end fashion match or improve results obtained through ad-hoc feature extraction procedures, achieving promising predictive accuracy without any input from domain experts.

The primary contributions of this work are:

- Using multimodal physiological time-series data from real-world and simulated driving environments to develop a stress recognition model with end-to-end representation learning on the one hand and manual feature engineering on the other.
- Demonstration of an unsupervised model adaption for cross-domain transfer using deep reconstruction classification networks.
- Presenting a robust approach for personalizing a model with deep multitask neural networks.

II. Model

A. Problem Definition

The stress detection (classification or recognition) can be framed as a sequence (time-series) classification task which takes physiological signals as input and outputs a label (generally binary) for each sequence. The raw input signals of different modalities are divided into segments of fixed length; with sliding window to avoid semantic segmentation. This process produces m input-output $\{(x_i, y_i)\}_{i=1}^{m}$ pairs, where y_i is taken to be the mode of context window. The $\vec{x_i}$ is either used directly for learning representations with deep networks or high-level features are extracted from it manually to learn a classification model.

B. Unsupervised Model Adaption

We formulate model adaption as a cross-domain and cross-user transfer learning problem. Here, a model trained on a dataset collected in a specific setting or source domain has to be adapted to perform the same task in a different situation or target domain. The key challenges, in this case, are a) unavailability of ground-truth for the target domain, b) expensive process of acquiring a large number of labels, and c) dynamic shift in data distribution. Therefore, target data cannot be directly used for fine-tuning an existing model in a supervised manner. However, the unlabeled target data provide auxiliary training information that can be leveraged to improve model generalization on the target domain than using only source data. This learning setting resembles MTL in the sense that learning an auxiliary task can help improve performance for the actual task using a shared representation [21].

Our goal is to transfer knowledge from labeled source data S to improve classification performance on unlabeled target data T. Let, X_S represent data instances and let Y_S be stress labels for the source. Likewise, X_T denotes data points from the target without any labels, Y_T. In domain adaption case, the marginal probability distribution of input data i.e. $P(X_S)$ and $P(X_T)$ are different but the set of classes are the same $Y_S = Y_T$. We used an extension of deep reconstruction classification network [20] to jointly model distribution of S and T with a combination of supervised and unsupervised objectives. The model is based on temporal convolution and recurrent layers (see Fig. 1). There are two distinct stages of the source and target feature learning by having a shared encoding representation. The initial stage is a hybrid of convolution and recurrent layers for source label predictions i.e. $\mathcal{C} : X_S \rightarrow Y_S$. While the subsequent phase is a denoising convolutional autoencoder for target data reconstruction i.e. $\mathcal{R} : X_T \rightarrow X_T$.

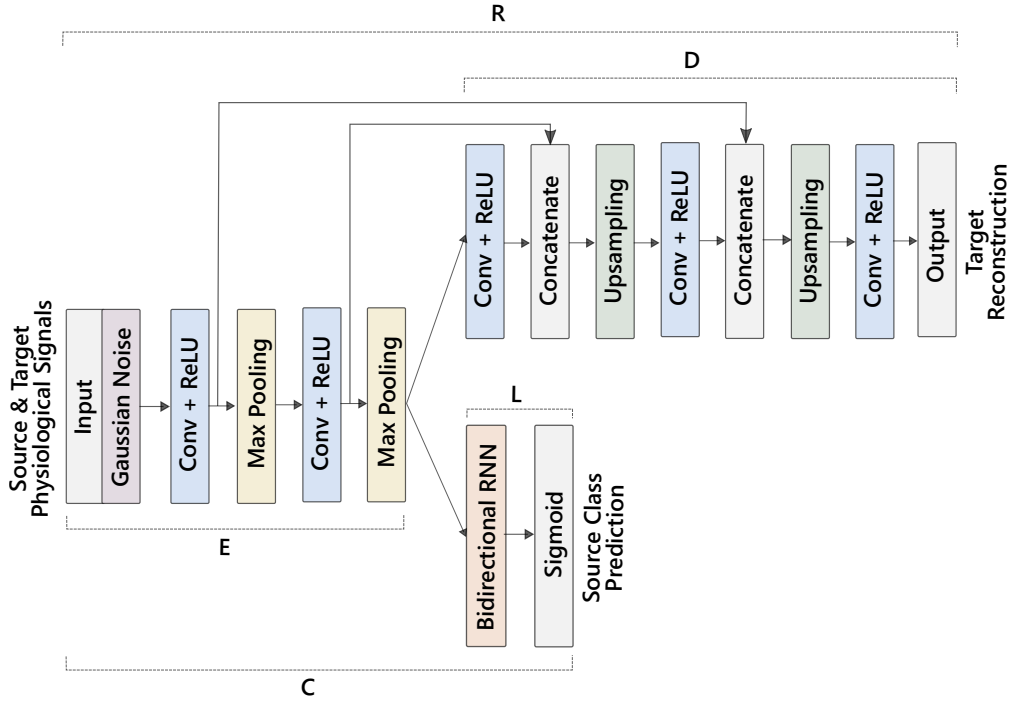

Fig. 1: Unsupervised (cross-domain) model adaption architecture. The network consists of three main blocks, encoder, decoder and label classifier, where encoder is shared between autoencoder and label classifier. The target data is reconstructed with encoder/decoder part of the network, represented by E and D. Similarly, source labels are predicted with the encoder and classifier, showed by E and L. The model is trained end-to-end with back-propagation using gradient descent (Adam). During optimization, first the weights of classification network (C) are updated followed by weight optimization of autoencoder (R). Concretely, the labeled source data flow through lower part of the model whereas, the unlabeled target data passes through upper part of the network.

The encoding phase of the architecture consists of 2 temporal convolution layers each followed by a max pooling operation with a pool size of 5. The convolution layers all have 90 feature maps and a filter length of 10 with rectified linear activation. The decoder architecture is similar except that the output is upsampled at the same rate as the input is downsampled in the encoder. The classification network shares the same encoder but has an additional bidirectional recurrent layer with 80 units. It is followed by a standard sigmoid layer to get a binary output. The Gaussian noise with a standard deviation of 0.1 is added to both source and target instances and l2-regularization is applied on the encoder's weights. The model is jointly optimized with binary classification ($\mathcal{L}_{\mathcal{C}}$) and reconstruction ($\mathcal{L}_{\mathcal{R}}$) losses for S and T, respectively. Given m_S source labeled instances $\{(x_i, y_i)\}_{i=1}^{m_S}$ and m_T unlabeled target samples $\{(x_i)\}_{i=1}^{m_T}$, the objective functions are then defined as follows:

$$\mathcal{L}_{\mathcal{CE}}(\hat{y}, y) = -\sum_{i=1}^{m}(y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})) \quad (1)$$

$$\mathcal{L}_{\mathcal{MSE}}(\hat{y}, y) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2 \quad (2)$$

$$\mathcal{L}_{\mathcal{C}} = \mathcal{L}_{\mathcal{CE}}(\mathcal{C}(X_S; \{\Theta_E, \Theta_L\}), Y_S) \quad (3)$$

$$\mathcal{L}_{\mathcal{R}} = \mathcal{L}_{\mathcal{MSE}}(\mathcal{R}(X_T; \{\Theta_E, \Theta_D\}), X_T) \quad (4)$$

$$\mathcal{L}_{\mathcal{CL}} = \alpha\mathcal{L}_{\mathcal{C}} + (1 - \alpha)\mathcal{L}_{\mathcal{R}} \quad (5)$$

where Θ_E, Θ_D, Θ_L, are an encoder, decoder and label prediction network weights, respectively. Note that Θ_E is shared between classification network $\mathcal{C}$ and autoencoder $\mathcal{R}$. Likewise, $0 \leq \alpha \leq 1$ is a trade-off hyper-parameter to control the contribution of classification and reconstruction losses.

C. Personalization

A subject-independent global model for stress detection may perform poorly due to large interpersonal variations in physiological parameters [19] e.g. due to age, gender, sleep, and diet. In order to take these disparities into account, we personalize a model by applying deep multi-task learning (MTL) with the subjects-as-tasks approach [22]. MTL involves finding a unified model for solving more than one task with a shared representation of the tasks. Consequently, a multi-task neural network (MT-NN) consists of common layers across tasks as well as task-specific layers. Besides, the last layer contains a separate output unit and a loss function for each task. The optimization of loss functions is done at the same time by alternating between different tasks at random.

We use two model architectures for the MTL setting, one based on the temporal convolutional neural network for

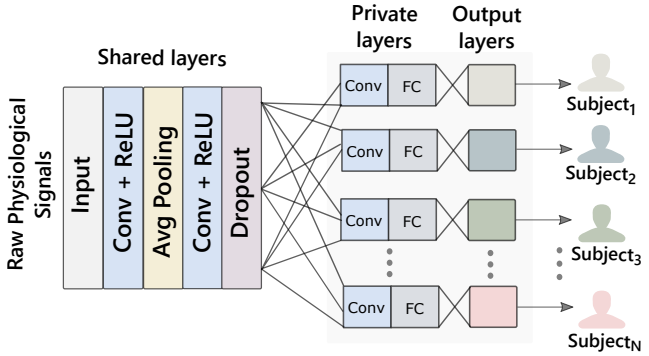

Fig. 2: Multi-task convolutional neural network architecture. The network consists of two temporal convolution and an average pooling layers along with a dropout which acts as the shared feature extractors. The separate convolution and dense layers are used as private layers for each participant with a sigmoid unit in the last. The model input is a 3d tensor of raw physiological signals of fixed length. It is trained end-to-end with back-propagation and gradient descent by alternating between tasks (subjects) at random (see Sec. III-D).

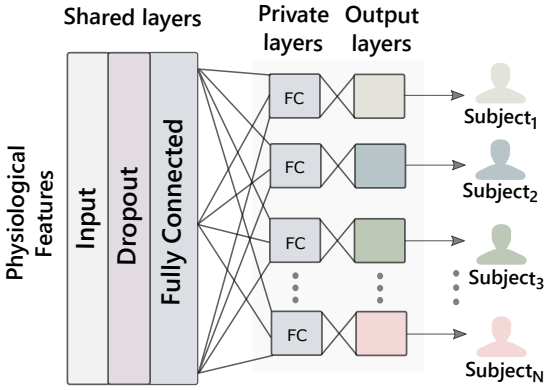

Fig. 3: Multi-task neural network architecture. The model consists of one shared layer (with hard-parameter sharing) and a (private) user specific dense layer with sigmoid classifiers in the last layer. The input is a vector of 16 physiological features, extracted from the heart rate and skin conductance manually (see Sec. III-D).

end-to-end representation learning (see Fig. 2) and another feed-forward neural network trained with manually extracted features. In the former, the first three layers act as the shared feature extractors among the tasks (see Fig. 3). They have 96 features maps with a kernel size of 8 and average pooling of size 5. A separate convolution and fully connected layers are employed as subject-specific layers to learn personalized features. The private convolution layer has the same configuration as shared ones but the dense layer has 512 hidden units with $tanh$ activation. In the latter, a fully-connected layer with 200 units is used as a shared layer, whereas a separate hidden layer with 100 units is used as a private layer for each subject. In both cases, the last part contains a sigmoidal layer with standard binary cross-entropy loss function (see Eq. 1) for each user. We use rectified linear activation in every layer (unless mentioned otherwise) and apply l2-regularization and dropout with a rate of 0.0001 and 0.2, respectively, to avoid over-fitting.

This model architecture will be able to take interpersonal variations in physiological signals into account through person-specific layers, rather than; having a mutual global representation. Likewise, we perceive personalization for new unseen user straight-forward through adding randomly initialized layers to an existing model. In this case, our architecture can be seen as an instantiation of Progressive Neural Networks [23]. The newly-added layers can be attached to existing shared layers; while dropping or chaining the private layers for knowledge transfer. The user-specific layers can then be optimized while keeping weights of shared layers frozen or tuning them separately with very small learning rate. This training strategy provides an additional benefit as the data from earlier users/tasks are not necessarily required to train a personalized model from scratch.

III. Dataset and Pre-Processing

We use skin conductance and heart rate signals from real-world and simulator driving datasets. The details are discussed below:

A. MIT Driver Stress (M)

The MIT Driver Stress dataset [24] consists of physiological signals recorded from 17 drives in a real-life experiment; when participants drove in a city, on a highway and rested in a garage. The collected signals comprise of EMG, ECG, galvanic skin response (GSR) from hand and foot, HR derived from ECG and respiration rate. The signals provided in the dataset are all down-sampled to 15.5 Hz. We used the 'marker' signal (a button press) to derive the ground-truth annotation for binary stress levels. The peaks are detected in the signal to capture the button push event; indicating a new trial of the experiment is commencing, e.g. the start or end of a rest period. The data points before and after the first and last markers (peaks) are removed as they correspond to the time when subjects were equipped with sensors. Likewise, 4 minutes of data after resting and before the beginning of the post driving baseline are removed. These steps are taken to avoid feeding signals with ambiguous labels, as it is hard to determine if subjects are stressed or recuperated. The artifacts are removed from HR and GSR signals following [25] as values fluctuated to unreasonably high and low levels. Likewise, ECG, GSR from foot and respiration rate are not used as collecting them in real-world situations is very problematic. Lastly, the following 10 drives dataset having valid HR, GSR from hand, and marker signals are used for further experimentation: 04, 05, 06, 07, 08, 09, 10, 11, 12 and 16.

B. Distracted Driving (D)

The multimodal Distracted Driving dataset [26] is acquired on a simulator in a controlled environment. The dataset includes data from 68 volunteers (35 male/33 female) that drove the same highway under four different conditions: a) no

interruption, b) cognitive distraction, c) emotional distraction, and d) sensorimotor distraction. In addition to the driving indicators (such as speed, brake force, and steering) and eye tracking; several physiological signals were recorded. These include palm electrodermal activity (EDA), HR, breathing rate, and perinasal perspiratory signal. We normalize EDA (dividing by 1000) as a pre-processing step to ensure the same range of variability compared to other data sources. In this research as our focus is on detecting cognitive load stressors, we used only the EDA and HR data (provided with a sampling rate of 1 Hz) from drive under normal and cognitive mental load. During a cognitive load drive, the stressors were induced by mathematical and analytical questions posed verbally by the experimenter. We used 40 participants for our analysis and dropped the rest due to corrupted or missing signals either during a normal or a stressful drive.

C. Cognitive Load Driving (C)

We collected heart rate and skin conductance (SC) data from 19 professional truck drivers using wrist-worn devices. The SC signal was recorded at a frequency of 10 Hz and HR was derived from Photoplethysmogram sensor data with a frequency of 1 Hz; it was upsampled to match the frequency of SC. The experiment was realized with a driving simulation software and participants received standardized instructions from an audiotape. The study consisted of three major steps 1) baseline driving, 2) moderate stress activity, and 3) high-stress task. The high stress was induced by means of a secondary arithmetic subtraction task. It is a component of widely used Trier Social Stress Test [27], where a user has to perform a serial subtraction verbally in a loud manner and has to start over from the last correct answer; if a mistake is made. Since we are interested in recognition of baseline and high stress, data points of moderate stress activity are dropped. Also, two subjects are dropped due to having bad quality signals.

D. Segmentation and Features

To prepare the data for model input, we used a sliding window approach as mentioned earlier to extract fixed-length sequences from each participant's physiological signals. A window length of 300 samples with a fixed step size of 50 samples is used for each dataset. In the case of end-to-end representation learning, raw physiological signals are used. For traditional learning algorithms, features are extracted manually from HR and SC which is discussed below. It is important to note that raw segments and features were computed from pre-processed signals, standardized with mean normalization by baseline to compensate for individuals having different resting heart rates.

Heart Rate: Heart rate is the number of complete cardiac cycles for instance measured as the R-R interval in an electro-cardiogram. It reflects the heart activity, including autonomic nervous system activity when it accommodates the body's demands depending on the received stimuli [10]. We obtained the following seven features from heart rate: mean, standard deviation, min, max, range, root mean square of successive differences, and standard deviation of successive differences.

Skin Conductance: The skin conductance (also known as galvanic skin response or electrodermal activity) describes the autonomic variations in electrical properties of the skin or equivalently, the number of active sweat glands. It is widely used as a sensitive index of emotional processing, sympathetic activity and is a relevant indicator of the stress level of a person [28; 29]. From this signal, the following nine features are extracted: mean, standard deviation, min, max, range, number of peaks, amplitude, skewness, and kurtosis.

IV. Experiments

Our experiments were conducted using physiological signals from three datasets described in Section III: MIT Driver Stress (M) [24], Distracted Driving (D) [26] and Cognitive Load Driving (C). The data of every subject in each dataset is randomly divided into training, validation and test sets of size 70%, 10%, and 20%, respectively. For each experiment, the networks are trained from scratch, initializing the weights with the *Xavier* technique [30]. We use the Adam [31] optimizer with the default parameters but used the validation set to find optimal learning rate and trade-off parameters (α). The optimal values of α are found to be between [0.2-0.7]. Finally, we employ validation based early stopping during the optimization process to further avoid over-fitting and improve the stress recognition rate.

We evaluated DRCN for model adaption on six combinations (source S $\rightarrow$ target T) of the above mentioned datasets: C $\rightarrow$ M, C $\rightarrow$ D, D $\rightarrow$ C, D $\rightarrow$ M, M $\rightarrow$ D and M $\rightarrow$ C and report kappa and area under the receiver operating curve (AUROC) on the held-out test set. For a baseline, we used a CNN model trained only on source data with architecture similar to the encoder part of the model as discussed in Sec. II-B. Furthermore, we also experimented with feed-forward networks trained with manually extracted features for both DRCN and source-only settings. The feed-forward models consist of 3 hidden layers with 128, 64 and 32 units with $tanh$ activation, where the decoder network has a similar configuration but layers in opposite order to reconstruct the original input vector of 16 dimensions. The results are summarized in Table I. The DRCN model trained end-to-end demonstrates a strong performance boost for the unsupervised cross-domain transfer learning problem. It achieves kappa of above 0.7 from the simulator to on-road and vice-versa from source-only baseline kappa of 0.6. It is important to note that, we used a fixed architecture for all six combinations of model adaption tasks to show predictive performance increase via joint training on source and target. We believe further improvement can be achieved if architectural components (e.g. number of kernels, kernel size, activation) are optimized for each adaption task separately. Likewise, convolutional models trained end-to-end outperformed those with an ad-hoc feature extraction procedure. This can be due to CNN's capacity and ability to automatically learn general to specific features from source and target domains together. Although, when

TABLE I: Test set Kappa and AUROC score for unsupervised (cross-domain) model adaption

	Methods	C → D	C → M	D → C	D → M	M → C	M → D
Kappa	Source Only - NN	0.040	0.640	0.371	0.604	0.470	0.148
	Source Only - CNN	0.246	0.648	0.389	0.566	0.594	0.401
	DRCN - NN	0.110	0.215	0.527	0.192	0.491	0.186
	DRCN - CNN	**0.541**	**0.656**	**0.747**	**0.701**	**0.774**	**0.432**
	Methods	C → D	C → M	D → C	D → M	M → C	M → D
AUROC	Source Only - NN	0.521	0.826	0.760	0.781	0.780	0.575
	Source Only - CNN	0.619	0.827	0.765	0.755	0.827	0.701
	DRCN - NN	0.563	0.628	0.807	0.610	0.798	0.598
	DRCN - CNN	**0.772**	**0.830**	**0.844**	**0.831**	**0.867**	**0.713**

the target domain is Distracted Driving, the domain adaption performance is comparatively low. This could be due to the relatively small size of this dataset and the recognition rate can be improved with a larger dataset.

In our attempt to personalize the model, we first evaluated two standard classifiers as a baseline: logistic regression (LR) and support vector machine with linear (L) and radial basis function (RBF) kernels. In addition, we also trained two layers (subject independent) feed-forward neural network with 100 hidden units and rectified linear activation in each layer. The data of each subject is randomly divided into (80/20) train and test sets. The cross-validation is performed on the training set for hyper-parameter optimization and evaluation metrics are averaged across participants on the test set. The stress recognition performance of these models is summarized in Table II, III and IV for real-world and simulator drivings, respectively. In MIT Driver Stress (on-road) dataset, SVM (RBF) set a strong baseline by achieving the highest results among other single-task models including ST-NN. The proposed MT-CNN model greatly improved upon that by achieving kappa of 0.84 and AUROC score of 0.91. It can be seen as an overall improvement across drivers due to subject-specific layers. Likewise, the MT-NN model which is trained with manually extracted features achieved similar results. Nevertheless, we advise caution in the interpretation of MIT Driver Stress dataset's result as no actual ground truth annotations or subjective self-reports are publicly available. The labels were acquired by means of a 'marker' signal, representing the start of next study trial (i.e. from resting to driving in a city) and assuming that driving, in general, is a stressful task.

For simulator driving datasets, the standard (one-fits-all) classifiers do not generalize as can be seen from the high standard deviation values of evaluation metrics in Table III and IV. The difference is particularly high for the Distracted Driving dataset, where a number of participants were comparatively large and more diverse belonging to different gender and age groups. The MT-NN notably improved the recognition rate across subjects and resulted in a better model by achieving kappa of 0.91 and 0.81 on both simulator datasets. Similarly, MT-CNN performed well apart from on Distracted Driving dataset which can be attributed to its small size as deep models require large datasets for representation learning. However, these results show that multi-task learning with reliable quality signals can be used to develop a personalized model as it generalizes well across various users and different environments i.e. real-world and simulators.

TABLE II: Average test set (20%) results of drives in MIT Driver Stress dataset

Model	AUROC	Kappa
LR	0.821 ± 0.074	0.650 ± 0.143
SVM (L)	0.832 ± 0.076	0.675 ± 0.146
SVM (RBF)	0.894 ± 0.035	0.808 ± 0.062
ST-NN	0.852 ± 0.116	0.707 ± 0.241
MT-NN	0.905 ± 0.056	0.831 ± 0.098
MT-CNN	$\mathbf{0.918 \pm 0.058}$	$\mathbf{0.841 \pm 0.110}$

TABLE III: Average test set (20%) results of participants of Cognitive Load Driving dataset

Model	AUROC	Kappa
LR	0.880 ± 0.161	0.745 ± 0.314
SVM (L)	0.876 ± 0.141	0.740 ± 0.264
SVM (RBF)	0.923 ± 0.104	0.853 ± 0.252
ST-NN	0.935 ± 0.072	0.855 ± 0.142
MT-NN	$\mathbf{0.960 \pm 0.056}$	$\mathbf{0.911 \pm 0.114}$
MT-CNN	0.956 ± 0.080	0.918 ± 0.147

TABLE IV: Average test set (20%) results of users in Distracted Driving dataset

Model	AUROC	Kappa
LR	0.734 ± 0.215	0.473 ± 0.431
SVM (L)	0.735 ± 0.217	0.472 ± 0.429
SVM (RBF)	0.882 ± 0.152	0.760 ± 0.909
ST-NN	0.860 ± 0.166	0.715 ± 0.334
MT-NN	$\mathbf{0.908 \pm 0.140}$	$\mathbf{0.814 \pm 0.282}$
MT-CNN	0.871 ± 0.127	0.738 ± 0.257

V. Conclusion

In this work, we proposed a solution for unsupervised cross-domain adaption and personalization of physiological stress recognition models with deep multi-task learning (MTL). The traditional learning approaches used for stress detection mostly (see [17] for a review) rely on sensor data (such as EMG, respiration rate, facial expressions and pupil dilation) that are very hard to acquire in a real-life situation to develop practical applications. Likewise, they do not explicitly address issues of end-to-end representation learning, covariate shift, and domain adaption. Therefore, these methods may perform

poorly when data distribution (of a source domain) training instances differs from test instances (of a target domain). Similarly, global subject-independent models do not generalize well to new test subjects because of large interpersonal variations in physiological parameters of individuals which can be due to age, gender, sleep, and diet. We used skin conductance and heart rate from real-world and simulator driving tasks to show: a) how models can be adapted to improve predictive performance on target domain in an unsupervised manner with deep reconstruction and classification networks (DRCN) and b) how to utilize multi-task learning (with subjects-as-tasks) to get personalized stress models. In our experiments, we found that the convolutional neural network based DRCN model outperforms the models trained only on source data and feed-forward networks utilizing manually extracted features. Likewise, in model personalization experiments, the MTL networks either trained end-to-end or with feature extraction procedures significantly improve the recognition rate across all datasets as compared to single-task models. We believe, if a wearable device provides reliable and high-quality signals, a real-time stress detection application can be developed to improve safety and well-being. In addition to stress classification in a driving environment, a future study may involve applying and investigating the performance of these methods in a daily-life context by comparing the model's outputs against subjective self-reports.

ACKNOWLEDGEMENT

SCOTT (www.scott-project.eu) has received funding from the Electronic Component Systems for European Leadership Joint Undertaking under grant agreement No 737422. This Joint Undertaking receives support from the European Unions Horizon 2020 research and innovation programme and Austria, Spain, Finland, Ireland, Sweden, Germany, Poland, Portugal, Netherlands, Belgium, Norway.

REFERENCES

[1] N. Schneiderman, G. Ironson, and S. D. Siegel, "Stress and health: psychological, behavioral, and biological determinants," *Annu. Rev. Clin. Psychol.*, vol. 1, pp. 607–628, 2005.

[2] T. G. Pickering, "Mental stress as a causal factor in the development of hypertension and cardiovascular disease," *Current hypertension reports*, vol. 3, no. 3, pp. 249–254, 2001.

[3] L. Goette, S. Bendahan, J. Thoresen, F. Hollis, and C. Sandi, "Stress pulls us apart: Anxiety leads to differences in competitive confidence under stress," *Psychoneuroendocrinology*, vol. 54, pp. 115–123, 2015.

[4] S. D. Holmes, D. S. Krantz, H. Rogers, J. Gottdiener, and R. J. Contrada, "Mental stress and coronary artery disease: a multidisciplinary guide," *Progress in cardiovascular diseases*, vol. 49, no. 2, pp. 106–122, 2006.

[5] A. L. Dougall and A. Baum, "Stress, health, and illness," *Handbook of health psychology*, pp. 321–337, 2001.

[6] J. Hernandez, R. R. Morris, and R. W. Picard, "Call center stress recognition with person-specific models," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 125–134.

[7] J. T. Cacioppo, L. G. Tassinary, and G. Berntson, *Handbook of psychophysiology*. Cambridge University Press, 2007.

[8] C. W. Sem-Jacobsen, "Electroencephalographic study of pilot stresses in flight," Gaustad Hospital Oslo (Norway) EEG Research Lab, Tech. Rep., 1961.

[9] A. Saeed and S. Trajanovski, "Personalized driver stress detection with multi-task neural networks using physiological signals," *NIPS workshop on Machine Learning for Health (ML4H), 8 December, 2017, Long Beach, CA, USA.*, 2017.

[10] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 2, pp. 156–166, 2005.

[11] D. A. Hennessy and D. L. Wiesenthal, "Traffic congestion, driver stress, and driver aggression," *Aggressive behavior*, vol. 25, no. 6, pp. 409–423, 1999.

[12] J. Zhai and A. Barreto, "Stress detection in computer users based on digital signal processing of noninvasive physiological variables," in *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*. IEEE, 2006, pp. 1355–1358.

[13] J. B. Sexton, E. J. Thomas, and R. L. Helmreich, "Error, stress, and teamwork in medicine and aviation: cross sectional surveys," *Bmj*, vol. 320, no. 7237, pp. 745–749, 2000.

[14] R. Zangróniz, A. Martínez-Rodrigo, J. M. Pastor, M. T. López, and A. Fernández-Caballero, "Electrodermal activity sensor for classification of calm/distress condition," *Sensors*, vol. 17, no. 10, p. 2324, 2017.

[15] T. Salafi and J. Kah, "Design of unobtrusive wearable mental stress monitoring device using physiological sensor," in *7th WACBE World Congress on Bioengineering 2015*. Springer, 2015, pp. 11–14.

[16] J. Wijsman, B. Grundlehner, H. Liu, J. Penders, and H. Hermens, "Wearable physiological sensors reflect mental stress state in office-like situations," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 600–605.

[17] N. Sharma and T. Gedeon, "Objective measures, sensors and computational techniques for stress recognition and classification: A survey," *Computer methods and programs in biomedicine*, vol. 108, no. 3, pp. 1287–1301, 2012.

[18] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1521–1528.

[19] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine

emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[20] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *European Conference on Computer Vision*. Springer, 2016, pp. 597–613.

[21] R. Caruana, "Multitask learning," in *Learning to learn*. Springer, 1998, pp. 95–133.

[22] N. Jaques, S. Taylor, E. Nosakhare, A. Sano, and R. Picard, "Multi-task learning for predicting health, stress, and happiness," in *NIPS Workshop on Machine Learning for Healthcare*, 2016.

[23] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.

[24] J. Healey and R. W. Picard, "Driver stress data," *Retrieved June 26th from MIT Affective Computing Group: http://affect.media.mit.edu*, vol. 124, 2002.

[25] S. Ollander, "Wearable sensor data fusion for human stress estimation," 2015, master Thesis, Technical University of Linkping University.

[26] S. Taamneh, P. Tsiamyrtzis, M. Dcosta, P. Buddharaju, A. Khatri, M. Manser, T. Ferris, R. Wunderlich, and I. Pavlidis, "A multimodal dataset for various forms of distracted driving," *Scientific data*, vol. 4, p. 170110, 2017.

[27] M. A. Birkett, "The trier social stress test protocol for inducing psychological stress," *Journal of visualized experiments: JoVE*, no. 56, 2011.

[28] E. Labbé, N. Schmidt, J. Babin, and M. Pharr, "Coping with stress: the effectiveness of different types of music," *Applied psychophysiology and biofeedback*, vol. 32, no. 3-4, pp. 163–168, 2007.

[29] P. Ferreira, P. Sanches, K. Höök, and T. Jaensson, "License to chill!: how to empower users to cope with stress," in *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*. ACM, 2008, pp. 123–132.

[30] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.