Multi-target Affect Detection in the Wild: An Exploratory Study

Philip Schmidt[*]
Robert Dürichen
Attila Reiss
Robert Bosch GmbH,
Corporate Research, Germany
first.last@de.bosch.com

Kristof Van Laerhoven
Ubiquitous Computing,
University of Siegen
Siegen, Germany
kvl@eti.uni-siegen.de

Thomas Plötz
School of Interactive Computing,
Georgia Institute of Technology
Atlanta, GA, USA
thomas.ploetz@gatech.edu

ABSTRACT

Affective computing aims to detect a person's affective state (e.g. emotion) based on observables. The link between affective states and biophysical data, collected in lab settings, has been established successfully. However, the number of realistic studies targeting affect detection in the wild is still limited. In this paper we present an exploratory field study, using physiological data of 11 healthy subjects. We aim to classify arousal, State-Trait Anxiety Inventory (STAI), stress, and valence self-reports, utilizing feature-based and convolutional neural network (CNN) methods. In addition, we extend the CNNs to multi-task CNNs, classifying all labels of interest simultaneously. Comparing the F_1 score averaged over the different tasks and classifiers the CNNs reach an 1.8% higher score than the classical methods. However, the F_1 scores barely exceed 45%. In the light of these results, we discuss pitfalls and challenges for physiology-based affective computing in the wild.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing**.

KEYWORDS

Affective Computing; User Study; Neural Networks

ACM Reference Format:
Philip Schmidt, Robert Dürichen, Attila Reiss, Kristof Van Laerhoven, and Thomas Plötz. 2019. Multi-target Affect Detection in the Wild: An Exploratory Study. In *Proceedings of the 2019 International Symposium on Wearable Computers (ISWC '19), September 9–13, 2019, London, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3341163.3347741

1 INTRODUCTION

Experiments conducted by *Bower et al.* [2] indicate that human decision making and memorization are strongly linked to their affective state. Affective computing (AC) aims at detecting the affective state (e.g. emotion) of a person based on some sort of observables. Common modalities for this purpose are for instance physiological indicators [27], smartphone data [32], or videos [6]. Many users of commercial mobile devices are interested in automatically logging information related to their physical health, e.g. step counts, consumed/burned calories, heart rate. Recently, a first generation of commercial devices promising insights into personal mental health, by detecting stress[1] [2], entered the consumer market. Providing users with data-driven insights into their, especially negative, affective states could help to create more awareness and lead to an overall improvement of health.

In order to quantitatively describe affective states, dimensional approaches are widely used [12, 15]. Following Russell, the affective space is partitioned along two axes (valence-arousal) and each emotional state can be placed in the corresponding space [24]. The valence scale ranges from negative to positive, whereas the arousal axis represents the level of activeness (low to high energy) [12]. Following this framework, excitement, for instance, can be placed in the positive valence/high arousal part of the 2D valence-arousal space.

Over the past decades many AC studies have been performed in lab settings [5, 15, 22, 26]. However, a shift from lab to field has been observed [28] recently. Unfortunately most of these studies are rather small scaled. In addition, there is a lack of datasets and studies containing physiological data and multi-dimensional labels, acquired at a fine-granular level. Nevertheless, these field studies bring everyday life affect detection systems, capable of detecting more than stress,

[*] Also with University of Siegen.

[1] www.apple.com/apple-watch-series-4/health/

[2] https://buy.garmin.com/en-US/US/p/567813

ISWC '19, September 9–13, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6870-4/19/09…$15.00
https://doi.org/10.1145/3341163.3347741

one step closer to their realization. Potential applications of such systems are diverse. First, affect aware machines (e.g., personal agents or robots) could use the affect of the user as an input, adjusting their behaviour accordingly. Moreover, in psychiatric care, affect detection system can help to improve diagnostics. *Gruenerbl et al.* [9], for instance, used smartphone data to detect state changes in bipolar disorder patients. In addition, real-word affect detection would provide users insights into their affective states. This could have the potential to uncover hidden correlations between mood swings and external factors.

Wearables, like smartphones and watches, facilitate out-of-the-lab AC. This is due to three reasons: first, their passive, unobtrusive, and ubiquitous sensing capability, second, their computational power, and third, their broad acceptance by a large number of users. However, due to misplacement or motion artefacts, data acquired using wearables is often noisy. In addition, human physiology also differs between subjects, depending on age, sex, and physical condition [3]. These data related issues aside, obtaining a crisp ground truth (e.g., precise labels on the affective state) is difficult. During field studies the subjects have to report their affective states via ecological-momentary-assessments (EMAs). These self-reports are subjective, and might drift over time. In addition, subjects are less likely to report on negative affective states (social desirability bias [7]). Furthermore, users are not able to file EMAs during certain events (e.g., exams or job interviews). Hence, these periods are likely to remain unlabelled, although they correspond to strong emotional stimuli. Until now, the detection results published for wearable affect detection in the wild are limited and can be sometimes worse than a baseline naive guessing [12].

Working with time-series data, recognition tasks are often tackled following a common pipeline: preprocessing, segmentation, feature extraction, and then using a classifier to solve the detection task. Following this classical pipeline, a suitable representation, also called feature engineering is one challenge. The fusion of different sensor modalities, often also sampled at different frequencies, is another. The first challenge can either be solved by using features known from literature or by using classifiers (e.g., CNNs), which are able to extract descriptive features from the (raw) input data. In addition, CNNs offer multiple ways to fuse sensory data [20]. Furthermore, end-to-end trainable CNNs easily incorporate multi-task learning and unsupervised pre-training. Operating on the 3-axis gyroscope and 3-axes acceleration (ACC), CNNs have been introduced to the human activity recognition (HAR) domain recently, exhibiting outstanding results [10, 11, 20]. Considering the success of CNNs in HAR and observing the similarity of data used for HAR and AC, we investigate the performance of CNNs on different affect recognition tasks. In this paper we present a realistic field study on affect recognition, applying latest machine learning (ML) methods and utilizing wearable biophysical modalities. We make the following contributions:

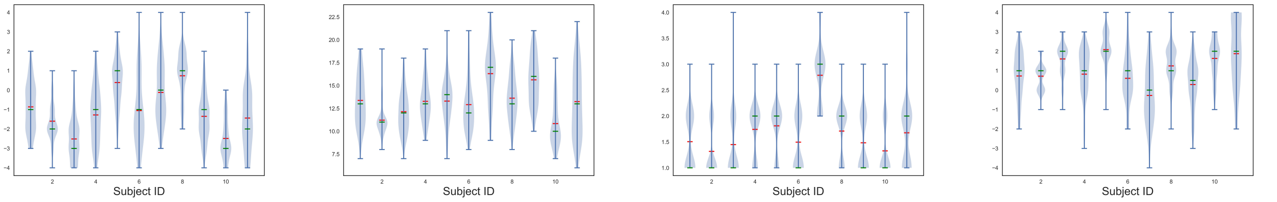
(1) We have recorded a novel realistic affect recognition dataset, containing multi-dimensional labels and multi-modal sensory data on 11 healthy subjects over more than two weeks on average.
(2) The data and labels acquired during this field study are used to formulate different affect recognition tasks. The performance of both classical (feature-based) ML methods and CNNs are compared, and we present a multi-input and multi-output, multi-task CNN (MT-CNN) architecture operating on the raw time series data predicting arousal, STAI, stress, and valence simultaneously.
(3) The classification scores for the classical baseline and the new CNN approaches are comparable to previous work. We provide a thorough discussion and analysis on the study's results, and list new insights on pitfalls posing limitations on affect recognition in the wild using wearable sensors.

2 RELATED WORK

In recent AC field studies, two major directions can be observed: first, stress detection and second, detection/forecasting of mood, using valence-arousal space [12] or happy/sad mood as targets [31].

Stress or workload detection in constrained environments, e.g., during driving tasks [13, 29] or at work [14], have been studied previously. *Hernandez et al.* [14] used electro-dermal activity (EDA) data, also known as skin conductance, of nine call centre employees to detect stressful calls. Comparing leave-one-subject-out (LOSO) and leave-one-day-out cross-validation the latter led to a 20% lower mean accuracy. This finding highlights the challenges of the subject independent evaluation. *Plarre et al.* [23] and *Gjoreski et al.* [8] present approaches where information from lab and field studies are combined to perform a context sensitive stress detection in the wild. In [8] the authors present a context-based stress detector and evaluate it on a small scale field study. The context for this stress recognition system is provided by an activity recognizer, a stress detector trained on lab data and further context information (e.g., hour, type of day - weekend/workday). Following this approach, a mean F1-score of 0.9 is reached. A major limitation of this study is its size (N=5) and, in addition, most of the data were provided by a single subject. Furthermore, the mean F1-score, reported in [8], drops to 0.47 if the detection is purely based on physiological and inertial data. This indicates that, although targeted by many, stress detection in the field is not yet a solved problem.

The topic of detecting affective states in the wild using dimensional representations ("high vs. low" arousal/"positive



(a) Arousal values. **(b) STAI values.** **(c) Stress values.** **(d) Valence values.**

Figure 1: Violin plots depicting the label distributions for each label type and subject. Mean in red. Median in green.

vs. negative" valence) has been studied by *Healey et al.* [12]. For this purpose a large dataset has been collected (~ 900 hours) recording EDA, activity, and cardiac data. However, the feature-based classifiers trained on this data and using 10-fold cross-validation, performed barely better than naive guessing. This clearly highlights the challenges for AC in the wild. In the domain of mood forecasting *Taylor et al.* [31] presented a multi-task approach which aimed to predict the mood, stress, and health of the following day. For this purpose the employed classifiers, different feature-based single and multi-task classifiers, e.g., multilayer perceptrons (MLPs), utilize physiological, context, and questionnaire-based features. The highest (binary) prediction accuracies (78%-82%) were reached employing a multi-task MLP and using groups of persons with similar personality traits (e.g., Big Five) as different sub-tasks. Although these results are supported by a large data base (N=104, recorded days 1842) this approach has a major drawback: it utilizes survey information (e.g., activities, social interaction, sleep) and passively acquired sensory data. This might be fine for the considered prediction task (tomorrow's values based on today's data). However, in a classification setup, detecting the *current affective state* based on current data, using survey data as input is not feasible.

Considering the body of related work, we aim to combine the detection of different affective states based on physiological data. As we are interested in real-time applications we limit ourselves to physiological and motion data only.

3 FIELD STUDY DATA COLLECTION

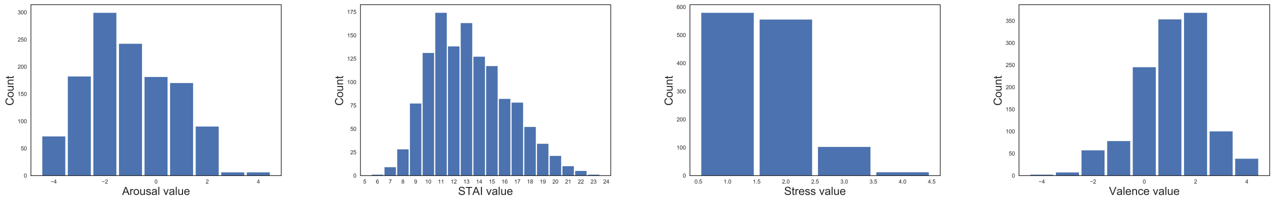
The quantitative analysis presented in this work is based on a field study dataset. While the study was still running, preliminary qualitative results, guidelines, and lessons learned have been presented in [27]. In total, 12 healthy subjects (7 male, 5 female) participated in the study. Due to sensor malfunction we had to exclude one participant (female). Hence, further analysis will be based on the remaining 11 subjects (mean age 26 ± 2.5, mean participation duration 16 ± 1 days). The study was approved by both the workers council and the data security officer of our research facility. The recruitment happened mostly via Email and all participants were students. During the study, physiological time series data, context information, and affective labels have been collected.

The physiological data was recorded using the Empatica E4 wristband, logging ACC (32 Hz), EDA (4 Hz), photoplethysmogram (PPG) (64 Hz), and skin-temperature (TEMP) (4 Hz). The study participants wore the E4 on their non-dominant hand during their wake hours. As a result, more than 1400 hours of physiological data have been recorded.

Prior to the study the subjects completed both a perceived stress scale (PSS) and a Pittsburgh Sleep Quality Index (PSQI) questionnaire. During the study labels were generated using a self-developed Android application (referred to as EMA App) running on the subjects' smartphone. During an initial face-to-face meeting the subjects were taught to use the EMA App. Furthermore, the EMA App was customised to the diurnal rhythm of each subject. During the configured time span (e.g., 7.30 to 22.30) the subjects received EMA prompts every 120 ± x minutes ($x \in [0, 30]$ and chosen randomly). In addition, the participants were instructed to manually trigger an EMA whenever they felt a change in their affective state. Daily screenings, conducted during the week, ensured a high data and label quality. Each EMA incorporated several types of questionnaires [27]. For this analysis the following subset of labels is considered:

- **Valence** and **arousal** labels generated using the well-known self-assessment mannequins [19].
- A shortened version (six items) of the State-Trait Anxiety Inventory (**STAI**) [30]. Items were chosen according to their factor loads, and scored on a four point Likert scale.
- **Stress** level scored on a four point Likert scale [8].

Using the subjects' smartphones and the EMA App context information was gathered too. This context information ranges from questionnaire data (e.g., sleep quality and intensity of last activity) to passively acquired information (e.g., weather, location, activity, and interaction data). However, we will limit our analysis presented below to the physiological data only. The reasons for this are twofold: first, the EMA



(a) Arousal Values. **(b) STAI values.** **(c) Stress values.** **(d) Valence values.**

Figure 2: Cumulated arousal, STAI, stress and valence values generated during the field study using our EMA App.

based context information should not be used in our opinion as it might not be available in most real world scenarios. Second, not all passively sensed data is available for all subjects, e.g., 2 out of 11 participants didn't agree to location logging.

Figure 1 displays violin plots of the arousal, STAI, stress, and valence label distributions for the different subjects. These plots highlight strong inter-subject differences: S4, S5, and S7, for instance, tend to be more stressed than the other participants. The histograms in Figure 2 depict the label distributions. Histograms and violin plots were generated using all filed EMAs. It becomes apparent that the corresponding label distributions are skewed: In general, the histograms exhibit only little mass in the high arousal, STAI, and stress bins. Judging form the valence scale, the subjects are more often in a positive than in a negative valence state.

Using Pearson's R, a correlation analysis has been performed for the labels. A strong negative correlation (-.60) between the STAI and valence values was found. Furthermore, a moderate positive correlation was found between the arousal and STAI values (.44) and a strong correlation is also observed between the stress and STAI values (.68). The above detailed correlations are significant (2-tailed p-values < 0.001). We found no correlation between valence and arousal labels. This finding emphasises that valence and arousal are independent scales.

4 METHODS

Windowing and questionnaire binning

Similar to [8] we labelled the time period from $X - 600$ sec to $X + 655$ sec with the affective states reported in the EMA started at time point X. The additional 55 seconds account for the completion of the entire set of questionnaires and was empirically verified prior to the data collection. In the next step, these time spans were segmented using a sliding window. Following, a recent review by *Kreibig*, the window size for the segmentation was set to 60 sec [16]. The window shift was 5 sec. Hence, for each valid time period 240 windows were extracted.

During the segmentation we excluded time periods where the E4 was either not worn or one of the sensors had a malfunction. Questionnaires with incomplete physiological data in the considered interval have been rejected, too. After the above detailed cleaning procedure a total of 1083 valid questionnaire were retained.

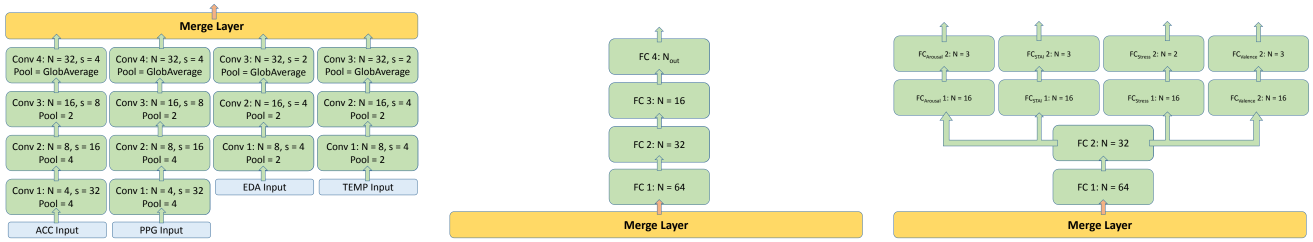
We formulated a three class classification problem for the arousal, STAI, and valence labels. The below detailed bins were chosen in order to establish equally sized bins for the three class classification tasks. Arousal was binned according to the following scheme: low (-5, -2], medium (-2, 1], and high (1, 5]. For valence the same bins were used, yielding a negative, neutral, and positive class. The employed STAI thresholds are (5, 12], (12, 18], and (18, 25]. Considering the skewness of the stress distribution (Figure 2c), a binary classification problem was formulated. For this purpose the labels with a value equal to 1 represent the "No stress" class. The other (2, 3, 4) were used to represent the "Stressed" class. In Table 1 the number of valid questionnaires per bin are displayed.

Evaluation Method and Metric

In order to validate our approaches leave-one-subject-out (LOSO) and leave-target-questionnaires-out (LTQO) were employed . For LTQO a stratified N-fold 80%/10%/10% (Train/ Test/ Validation) split was performed. The stratified nature of these splits ensures similar label distributions in the different splits. However, in contrast to simple N-fold cross-validation, this scheme ensures that all instances belonging to specific *target questionnaires* are placed in the same set. Hence, this validation scheme provides an insight into the subject depended performance of the different classifiers. This scheme

Table 1: Number of questionnaires in the different bins.

	Low	Medium	High
Arousal	479	519	85
STAI	479	539	65
	negative	neutral	positive
Valence	56	593	434
	No stress	Stressed	
Stress	504	579	-



(a) Feature extractor applying a four and three layered CNN architecture to the windowed data.

(b) Single-task CNN (ST-CNN) receive input from the feature extractor (Figure 3a).

(c) Multi-task CNN (MT-CNN) receive input from the feature extractor (Figure 3a).

Figure 3: Illustration of the building blocks of the CNNs. ST-CNN and MT-CNN are constructed by placing them on top of the feature extractor, Figure 3a.

has been employed to mitigate the large individual differences in the label distributions displayed in Figure 1.

Due to the skewness of the considered dataset the macro F_1 score, corresponding to the unweighted mean of the F_1 scores for the different labels, is used as evaluation metric. The performance of the different ML classifiers is compared to the performance of a sophisticated guesser (also known as Zero Rule). This classifier always predicts the majority class found in the training data. Later, an investigation of different types of classifiers (feature-based and end-to-end learning) is presented and their performance is compared.

Features and classical approach

For the classical experiments we followed the HAR pipeline [4] and extracted features from windowed data (size 60 sec, shift 5 sec). In total 62 features were extracted and used as input for the classical classifiers. We used the same set of E4 features as described in Table 1 by *Schmidt et al.* [26]. These features range from plain statistical features (mean and standard deviation), to complex physiological features like heart rate, heart rate variability, or number of peaks in EDA data. Here, two different experiments were performed: in the first experiment, the features were used directly as input. In the second, a z-transformation, normalizing each feature to zero mean and unit variance has been applied. For the classical evaluation the sklearn [21] implementation of different tree-based classifiers (decision-tree (DT), randomized decision trees (ET), and random-forest (RF)) have been used.
For ensembles (RF and ET) the number of trees were chosen to be N=101. In order to avoid overfitting, the minimal number of samples per split was set to 150 for all classical classifiers.

Table 2: Overview over the numbers of convolutional layers in the different feature extraction branches and the corresponding parameters. Abbreviations: fully-connected (FC), rectifying linear unit (RELU).

	ACC+PPG 32 Hz	EDA+TEMP 4 Hz
# Filter per Layer	4, 8, 16, 32	8, 16, 32
Kernel size	32, 16, 8, 4	4, 2
Max-pool	4, 4, 2	2, 2
Stride + Padding	1 + ’same’	1 + ’same’
Non-linearity	RELU	RELU
Neurons in FC layer	64, 32, 16, N_{out}	

End-to-End learning

In the end-to-end learning scenario the windowed data served as direct input into CNNs. Below we present an approach for affect recognition based purely on physiological time series data utilizing CNNs. Starting with the single-task CNN (ST-CNN) formulation we extend this approach to a multi-task CNN (MT-CNN) classifier, predicting arousal, STAI, stress, and valence simultaneously.

Convolutional Neural Networks

The ST-CNNs and MT-CNNs architectures investigated here utilize four layer types: convolutional, max-pooling, global-average pooling [17], and fully-connected (FC) layers. The CNNs receive the windowed E4 data (ACC, EDA, PPG, TEMP) as input. As the PPG data has been down sampled by a factor of two the CNNs deals with two sampling frequencies (4/32 Hz).

Feature Extraction: The CNNs employ sensor-based late fusion [20]. This enables the network to learn modality-specific filters. The feature extraction part of the CNNs is depicted in Figure 3a. The architectural parameters (e.g., kernel size, stride, etc.) were chosen to be the same in branches with the same sampling frequency (ACC+PPG and EDA+TEMP). In each of these branches convolution and max-pooling layers are alternated. Table 2 details the hyperparameters used in the feature extraction branches. Throughout the network

RELUs, are employed as non-linearities. After the feature extraction a global average pooling operation is performed.

Classification: Both the ST-CNN and MT-CNN approach utilize the feature extraction architecture described above (see Figure 3a). The main difference between the two approaches lies in the classification part of the network: the ST-CNN uses a separate feature extractor for each classification task and on top four FC layers (depicted in Figure 3b), classifying *one specific* type of label each (e.g., valence). In contrast, the MT-CNN share two FC layers, have multiple output branches (see Figure 3c), and are trained to classify *all* labels types (arousal, STAI, stress, and valence) simultaneously. Apart from the last FC layer where a softmax is used, the FC layers also use RELUs as non-linearities. In both cases the CNNs were trained using a cross-entropy loss and mini-batches of size 1024 or 64. Following the hyperparameter settings of [11], ADAM was used as optimizer.

Following LTQO the ST-CNNs were trained, validated, and tested on stratified splits (80%/10%/10%) of the target questionnaire. In the MT-CNNs setup a stratified split was performed along the arousal labels and the split for the arousal values was then utilized for the other questionnaires, too. During training the F_1 score on the validation set was monitored. For prediction the weights corresponding to the highest score were used.

The number of trainable parameters differs between the ST-CNN and MT-CNN. The binary ST-CNN stress detector has 21946 parameters. Each ST-CNN architecture employed for arousal, STAI, and valence classification contains 21963 trainable parameters. Hence, in the ST-CNN formulation predicting arousal, STAI, stress, and valence would require four different CNNs with a total of 87835 parameters. In contrast, a MT-CNN has a total of 23683 parameters and predicts all targets of interest simultaneously. This is a factor 3.7 less parameters than in the ST-CNN approach.

Unsupervised pre-training: Similar to *Zheng et al.* [33] unsupervised pre-training, using convolutional auto-encoders (conv. AE), has been investigated. Here conv. AEs were trained for each sensor modality separately, using ADAM as optimizer and mean squared error loss. All conv. AE were trained for 40 epochs, using the windowed data (80%/20% train/test). Apart from the global average pooling operation the modality specific encoder employed the same type of convolution and pooling operations as the feature extractors described above. In the decoder part of the conv. AE upscaling and convolutional layers (reversing the number of kernels and filter sizes) were applied. Using these conv. AE weights two different experiments were performed:

(1) The encoder weights were set to non-trainable during the fine-tuning (referred to as *frozen* below). Hence, only the final FC classification layers were updated during training.
(2) Both the encoder weights and the classification layers were updated during the training. This setup is referred to as *not frozen.*

Here ADAM, with hyperparameter settings as before, has been used. The CNNs were implemented in keras with a tensorflow backend and trained on Nvidia GTX 1080 TI GPUs.

5 RESULTS

Feature-based Evaluation

Using the above described features and evaluation schemes (leave-one-subject-out (LOSO) and leave-target-questionnaires-out (LTQO)), the performance of different classical decision tree-based classifiers has been investigated. The performance of these classifiers is compared to a sophisticated guesser baseline, predicting only the majority class found in the training set. Table 3 displays the F_1 scores generated using LOSO. The results were averaged over all subjects and per subject five runs have been performed. Here, the decision tree-based classifiers are able to outperform the sophisticated guessing baseline by a large margin. The results of both, the sophisticated guesser baseline and the decision tree-based classifiers, have rather large standard deviations. This is to be attributed to the large inter-subject differences in the label distributions, which pose limitations on successful generalization. Averaging the obtained F_1 scores over the different tasks, the ET using the normalised features (ET+zN) reached the highest combined F_1 score.

In Table 4 the performance of the feature-based classifiers using the LTQO evaluation scheme is reported. In this setup the RF and ET reach similar averaged F_1 scores. The reasons for the increased F_1 scores following LTQO are twofold: first, LTQO is subject dependent, which simplifies the problem. Secondly, due to the stratified split the folds have the same label distributions, which decreases the standard deviation. From both Table 3 and Table 4 two major observations can be made: first, the normalisation has no crucial influence on the averaged F_1 scores. Second, the overall $F1$ scores are not satisfying. However, related work reported an F_1 score of .47, using similar feature-based methods for a stress detection task [8]. Hence, it can be speculated that the employed classifiers might not be powerful enough to learn these relations. Therefore, the next section explores the performance of CNNs, which established the current state-of-the-art on many HAR tasks [10, 11, 20].

End-to-End learning

Table 5 displays the results from the CNN experiments. For the binary stress classification task the highest F_1 scores were reached. Similar to the classical results presented in Table 4, the lowest scores are obtained for the multi-class STAI

Table 3: Mean F_1 score using feature-based classifiers and LOSO. Results are averaged over the different subjects and five runs per subject. The last column displays the F_1 score averaged over the tasks.

	Arousal	STAI	Stress	Valence	Average
DT	30.9 ± 3.8	31.4 ± 3.4	46.7 ± 5.9	33.5 ± 1.7	35.6 ± 3.7
DT+zN	30.8 ± 3.9	31.3 ± 3.3	**46.8 ± 5.7**	33.7 ± 1.9	35.7 ± 3.7
ET	31.4 ± 8.1	33.3 ± 9.0	46.5 ± 7.4	41.1 ± 8.2	38.1 ± 8.2
ET+zN	**31.4 ± 8.1**	**33.6 ± 9.3**	46.5 ± 7.3	**42.2 ± 9.5**	**38.4 ± 8.6**
RF	30.6 ± 5.1	31.7 ± 7.6	46.2 ± 7.2	39.9 ± 8.3	37.1 ± 7.1
RF+zN	30.6 ± 5.2	32.1 ± 8.5	46.2 ± 7.3	39.9 ± 8.4	37.2 ± 7.3
Base	19.9 ± 9.5	21.3 ± 6.7	39.0 ± 20.3	26.4 ± 7.9	26.6 ± 11.1

Table 4: Mean F_1 score using feature-based classifiers and the subject dependent validation scheme (LTQO). The displayed results are averaged over five runs and the last column displays the F_1 score averaged over the tasks.

	Arousal	STAI	Stress	Valence	Average
DT	37.4 ± 0.8	34.9 ± 1.4	53.1 ± 0.7	40.1 ± 1.5	41.3 ± 1.1
DT+zN	38.7 ± 2.2	37.5 ± 1.0	52.3 ± 0.4	39.6 ± 1.4	42.0 ± 1.3
ET	38.4 ± 1.5	36.8 ± 1.7	56.3 ± 1.7	42.8 ± 3.1	43.6 ± 2.0
ET+zN	38.8 ± 1.7	35.2 ± 1.9	57.9 ± 3.2	43.3 ± 1.2	43.8 ± 2.0
RF	38.2 ± 2.2	37.4 ± 2.0	54.9 ± 1.6	42.2 ± 0.6	43.2 ± 1.6
RF+zN	38.0 ± 2.5	36.0 ± 2.0	55.8 ± 1.7	42.6 ± 1.6	43.1 ± 1.9
Base	21.7	22.2	34.9	23.6	25.6 ± 5.5

Table 5: Mean F_1 score using CNNs and LTQO as validations scheme. All results were averaged over three runs and the last column displays the F_1 score, averaged over the different tasks.

		Arousal	STAI	Stress	Valence	Average
$Batch_{size} = 1024$	Training CNNs from scratch					
	ST-CNN	44.3 ± 1.4	39.2 ± 0.5	55.5 ± 3.4	42.8 ± 2.4	45.4 ± 1.9
	MT-CNN	42.8 ± 3.8	37.4 ± 1.1	56.6 ± 1.6	44.0 ± 2.4	45.2 ± 2.2
	Fine-tuning: conv. AE weights are *frozen*					
	ST-CNN	43.1 ± 4.1	36.9 ± 0.7	58.3 ± 0.8	43.2 ± 2.0	45.4 ± 1.9
	MT-CNN	39.4 ± 1.8	36.3 ± 0.5	56.8 ± 1.5	41.2 ± 0.8	43.4 ± 1.1
	Fine-tuning: conv. AE weights are *not frozen*					
	ST-CNN	42.5 ± 3.3	38.1 ± 2.2	53.8 ± 6.0	40.4 ± 1.4	43.7 ± 1.9
	MT-CNN	43.9 ± 3.0	41.5 ± 2.0	55.7 ± 2.7	39.0 ± 0.5	45.0 ± 2.0
$Batch_{size} = 64$	Training CNNs from scratch					
	ST-CNN	42.9 ± 3.0	37.0 ± 0.9	56.9 ± 1.4	40.3 ± 0.9	44.3 ± 1.5
	MT-CNN	42.1 ± 1.0	38.3 ± 1.6	57.0 ± 1.1	44.6 ± 3.5	45.5 ± 1.8
	Fine-tuning: conv. AE weights are *frozen*					
	ST-CNN	40.5 ± 1.6	35.1 ± 0.7	54.2 ± 2.0	43.4 ± 3.6	43.3 ± 2.0
	MT-CNN	42.8 ± 5.9	36.6 ± 1.2	56.6 ± 0.9	42.0 ± 1.6	44.5 ± 2.4
	Fine-tuning: conv. AE weights are *not frozen*					
	ST-CNN	41.9 ± 3.9	38.8 ± 0.7	55.8 ± 1.2	41.7 ± 1.6	44.5 ± 1.8
	MT-CNN	40.6 ± 2.4	39.2 ± 3.3	57.7 ± 0.9	41.6 ± 0.9	44.8 ± 1.9

classification. Using CNNs the arousal and valence classification tasks are solved with a similar performance. However, comparing classical valence and arousal classification (see Table 4) to the CNN-based one, especially the arousal task, is solved with a higher F_1 score.

The mean $F1$ score over the different tasks using CNNs is on average 1.8% better than the average $F1$ score of the feature-based classifiers. In general, this is only a minor improvement. The highest F_1 scores reached by the CNNs is around 45.5%. This result is achieved by the ST-CNN trained from scratch with $N_{batch} = 1024$ and the fine-tuned ST-CNN, where the weights of the convolutional auto-encoders (conv. AE) have been frozen. In addition, the MT-CNN trained from scratch setting $N_{batch} = 64$ reaches an averaged F_1 of 45.5%, too. Hence, these best CNNs outperform the best classical approach (ET, averaged $F_1 = 43.8$) by 1.6-1.7%.

In general, the performance of the ST-CNNs and MT-CNNs are comparable. However, the MT-CNNs predict all labels simultaneously and require only little more parameters than a single ST-CNN.

Judging from Table 5 both investigated batch sizes $N_{batch} = [1024, 64]$ led to similar performances. In addition, utilizing the pre-trained conv. AE weights did not improve the results. These observations hold for both experimental settings (frozen and not frozen weights in the feature extractor). The F_1 scores presented in Table 5 are only a marginal improvement of the scores reached by the classical approaches. However, the utilized CNNs operate directly on the windowed data and, hence, make feature engineering obsolete. In addition, the presented CNN are small (less than 25k parameters). Restricted Boltzmann Machines, requiring up to a factor of 140 more parameters, were successfully deployed on a Snapdragon 400 platform [1]. Therefore, the deployment of our models should be feasible on a similar platform.

6 DISCUSSION

Using data acquired during an affective computing (AC) field study, we investigated and compared the performance of feature-based classifies and convolutional neural networks (CNNs). Even for state-of-the-art CNNs it was challenging to reach an average F_1 score higher than 45%. From literature, it is known that the posed classification problem is very challenging [8, 12]. In our approach we see the following limitations:

Algorithmic: Our pipeline was tailored to directly classify

the affective state based on features, either learned or handcrafted, from 60 sec data snippets. In this approach the temporal and sequential nature of the data are not captured. This could be improved by adding a refinement step, e.g., voting over multiple adjacent windows. Another approach would be to model the temporal nature of the data explicitly by employing a Hidden Markow Model or LSTM, for instance.
Data and label quality: Data obtained from field studies are intrinsically noisy and the labels are not completely reliable. Data noise ranges from sensor misplacement to movement artefacts. Furthermore, label fuzziness can be attributed to the subjective nature of ecological-momentary-assessments (EMAs). We observed an intrinsic bias towards positive labels, this is reflected by the skewed label distributions in Figure 2. In our opinion the reasons for this skewness are twofold: first, the subjects are less likely to respond to or trigger an EMA, while being in a high arousal (e.g. stressed) affective state. Secondly, according to the social desirability bias [7] subjects are less likely to report on states less socially desired (like being in a bad mood). All in all, the data and label noise certainly has an adverse effect on the results.
Amount of labelled data: Labels gathered via EMAs are discrete and sparse. For the presented analysis we utilized 1083 valid EMAs. Training classifiers on such small amounts of (skewed) data is difficult and combating both over- and under-fitting is challenging, even if different types of regularization, e.g. dropout or L2 regularization, are employed.

Based on the analysis presented above we formulate lessons learned and identify pitfalls for AC in the wild:
Curse of normality: Healthy users are unlikely to exhibit strong mood swings across the entire affective spectrum. Assuming a (skewed) Gaussian shape of the label distribution most labels will be reported around a mean value of "things are okay/normal". As a result, extrema in the affective spectrum are broadly underrepresented. In the label data presented, see Figure 2, this "normal state" is indicated by low arousal/positive valence and other states are underrepresented. Hence, classifying rare episodes where a user is in an extreme state is, due to the low number of data points, challenging. However, treating extreme cases as outliers and applying methods from outlier detection is a direction worthwhile investigating. This could also be used during data collection to trigger EMAs, once an outlier state is detected.
Awareness of affective states: Between 5 and 18% of the general population has difficulties with identifying and describing their emotions [18]. Hence, label quality could be increased dramatically by providing mindfulness sessions for study participants. In addition, it might be interesting to explore other labelling techniques than EMAs, where subjects are given more time to reflect about their affective state and then answer a set of questionnaires.
Representation of affect: Dimensional representations (e.g., valence and arousal) of affective states are intuitive. Based on our study it seems, see Figure 1, that most subjects do not utilize the entire spectrum. This might be due to personal biases (e.g., personality traits). One way to mitigate this bias could be to normalize the labels of each subject. However, this approach would require the label distribution to be known. This is not the case, especially in a real world application. An alternative approach would be to develop affective scales with a finer granularity tailored to certain personality types. For instance, if someone claims to be a rather positive person it might be beneficial to inquire finer granular levels of positivity (e.g., 'less than normal', 'normal', 'more than normal'), instead of asking about negative valence. Another idea would be to ask the user to compare events (e.g., "Are you currently more/less aroused compared to your last report?"). Both approaches would increase the variance in the label distribution, facilitating the uncovering of hidden correlations.
Human activity recognition vs. affective computing: Both wearable-based HAR and AC utilize similar inputs to create a user model. In the HAR domain, however, the employed sensors, ACC for instance, offer direct measures of the performed activity (e.g., walking). In contrast, considering AC, the available sensors only offer indirect measures. In our opinion this contributes to the large performance gap between HAR and AC detection systems.
Modalities: In our evaluation, we aimed at classifying the affective state of a person purely based on physiological data. Although subjects cannot actively influence their physiological responses, there are many confounding variables. Judging from our experiments, the classifiers had difficulties identifying these confounders. Based on literature [8], context data might be able to alleviate this. However, another direction could also be to add more informative biomarkers to the picture, e.g., cortisol level. Furthermore, sleep quality has been proven to be a powerful predictor for mood [25]. Hence, this information could also help to improve the classification results. However, here the sleep quality information should be acquired in a passive fashion.
Discrete vs continuous: In contrast to biophysical signals like electro-dermal activity which is available continuously, other information like sleep quality or cortisol measures are only available once a day or at discrete time points. Combining both types of information in a single model can be challenging. One approach could be to have different models for different scenarios (e.g., one specific mood classifier for a high and one for a low cortisol level). Alternatively, these discrete values could serve as one feature used by the classifiers.

7 CONCLUSION

In this paper we presented approaches to classify the multi-label affective state of a person purely based on physiological time series data. The data was collected from 11 healthy subjects during a field study (mean participation 16 ± 1 days duration). Specifically, 3-axes acceleration, electro-dermal activity, photoplethysmogram, and skin-temperature were used to classify the arousal, STAI, stress, and valence self-reports. We employed both a subject independent, LOSO, and a subject dependent, LTQO, evaluation scheme. In a subject independent LOSO evaluation, feature-based classifiers reached average F_1 scores between 31% (three class arousal) and 47% (binary stress detection). In the subject dependent formulation, the performance of feature-based classifiers and different CNNs is compared. The CNNs lead, compared to the classical methods, to a minor improvement of the average F_1 score (1.8%).

Our results indicate that despite state-of-the-art methods AC in the wild is still very challenging. Hence, we discussed potential pitfalls for wearable-based AC and encourage the community to pick up on these challenges.

ACKNOWLEDGMENTS

We would like to thank Hyeokhyen Kwon for many fruitful discussions. We also thank all study participants for their commitment.

REFERENCES

[1] S. Bhattacharya and N. D. Lane. 2016. From smart to deep: Robust activity recognition on smartwatches using deep learning. In *PerCom Workshops*.

[2] Gordon H Bower. 1981. Mood and memory. *Am Psychol* (1981).

[3] B. Brenner, A.and Ehmke H. Burckhardt, G.and Draguhn, U. Eysel, and D. Singer. 2018. *Physiologie*.

[4] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput Surv.* (2014).

[5] J. Choi, B. Ahmed, and R. Gutierrez-Osuna. 2012. Development and Evaluation of an Ambulatory Stress Monitor Based on Wearable Sensors. *IEEE Trans. Inf. Technol. Biomed.* (2012).

[6] V. Ebrahimi Kahou, S.and Michalski, K. Konda, R. Memisevic, and C. Pal. 2015. Recurrent Neural Networks for Emotion Recognition in Video *(ICMI '15)*.

[7] A Edwards. 1957. The social desirability variable in personality assessment and research. (1957).

[8] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski. 2017. Monitoring stress with a wrist device using context. *J. Biomed. Inform.* (2017).

[9] A. Grünerbl, A. Muaremi, V. Osmani, G. Bahle, and P. Lukowicz. 2015. Smartphone-Based Recognition of States and State Changes in Bipolar Disorder Patients. *IEEE J BIOMED HEALTH* (2015).

[10] N. Hammerla, S. Halloran, and T. Ploetz. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables. *arXiv preprint arXiv:1604.08880* (2016).

[11] J. Hannink, T. Kautz, C. F. Pasluosta, K. Gaßmann, and B. M. Eskofier. 2017. Sensor-Based Gait Parameter Extraction With Deep Convolutional Neural Networks. *IEEE J. Biomed. Health Inform.* (2017).

[12] J. Healey, L. Nachman, S. Subramanian, J. Shahabdeen, and M. Morris. 2010. Out of the Lab and into the Fray: Towards Modeling Emotion in Everyday Life *(Pervasive'10)*.

[13] J. A. Healey and R. W. Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Trans. Intell. Transp. Syst.* (2005).

[14] J. Hernandez, R. Morris, and R. W. Picard. 2011. Call Center Stress Recognition with Person-Specific Models *(ACII'11)*.

[15] S. Koelstra, C. Muhl, M. Soleymani, J. Lee, T. Ebrahimi, and I. Patras. 2012. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput* (2012).

[16] Sylvia D Kreibig. 2010. Autonomic nervous system activity in emotion: A review. *Biol. Psychol.* (2010).

[17] M. Lin, Q. Chen, and S. Yan. 2013. Network In Network. *arXiv preprint arXiv:1312.4400*.

[18] A. Mattila, J. Salminen, T. Nummi, and M. Joukamaa. 2006. Age is strongly associated with alexithymia in the general population. *J Psychosom Res* (2006).

[19] Jon D Morris. 1995. Observations: SAM: the Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response. *J Advert Res* (1995).

[20] S. Münzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. Dürichen. 2017. CNN-based Sensor Fusion Techniques for Multimodal Human Activity Recognition *(ISWC '17)*.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, E. Duchesnay, et al. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12 (2011).

[22] R. Picard, E. Vyzas, and J. Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Trans. Pattern Anal. Mach. Intell.* (2001).

[23] K. Plarre, A. Raij, S. Hossain, M. Ali, A. Scott, et al. 2011. Continuous inference of psychological stress from sensory measurements collected in the natural environment *(IPSN 2011)*.

[24] James A Russell. 1979. Affective space is bipolar. *J Pers Soc Psychol* (1979).

[25] A. Sano, A. Yu, A. McHill, A. Phillips, and R. Picard. 2015. Prediction of Happy-Sad mood from daily behaviors and previous sleep history. In *Conf Proc IEEE Eng Med Biol Soc.*

[26] P. Schmidt, A. Reiss, R. Dürichen, C. Marberger, and K. Van Laerhoven. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection *(ICMI '18)*.

[27] P. Schmidt, A. Reiss, R. Dürichen, and K. Van Laerhoven. 2018. Labelling Affective States "in the Wild": Practical Guidelines and Lessons Learned *(UbiComp '18)*.

[28] P. Schmidt, A. Reiss, R. Dürichen, and K. Van Laerhoven. 2018. Wearable affect and stress recognition: A review. *arXiv preprint arXiv:1811.08854* (2018).

[29] S. Schneegass, B. Pfleging, N. Broy, F. Heinrich, and A. Schmidt. 2013. A data set of real world driving to assess driver workload *(AutomotiveUI '13)*.

[30] C. Spielberger, R. Gorsuch, and R. Lushene. 1983. Manual for the state-trait anxiety inventory. (1983).

[31] S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard. 2018. Personalized Multitask Learning for Predicting Tomorrows Mood, Stress, and Health. *IEEE Trans. Affect. Comput.* (2018).

[32] X. Zhang, W. Li, X. Chen, and S. Lu. 2018. MoodExplorer: Towards Compound Emotion Detection via Smartphone Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* (2018).

[33] Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J. Leon Zhao. 2016. Exploiting multi-channels deep convolutional neural networks for multivariate time series classification. *Front. Comput. Sci.* (2016).