An Inter-domain Study for Arousal Recognition from Physiological Signals

Martin Gjoreski[1,2], Mitja Luštrek[1] and Matjaž Gams[1,2]
[1]Department of Intelligent Systems, Jožef Stefan Institute
[2]Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
E-mail: martin.gjoreski@ijs.si

Blagoj Mitrevski
Faculty of Computer Science and Engineering
Skopje, R. Macedonia

Keywords: arousal recognition, GSR, R-R, machine learning, emotion recognition, health

Received: October 27, 2017

Arousal recognition from physiological signals is a task with many challenge remaining, especially when performed in several different domains. However, the need for emotional intelligent machines increases day by day, starting with timely detection and improved management of mental disorders in mobile health, all the way to enhancing user experience in human-computer interaction (HCI). One of the open research questions, which we analyze in this paper, is which machine-learning (ML) methods and which input is most suitable for arousal recognition. We present an inter-domain study for arousal recognition on six different datasets. The datasets are processed and translated into a common spectro-temporal space of R-R intervals and Galvanic Skin Response (GSR) data, from which features are extracted and fed into ML algorithms. We present a comparison between dataset-specific models, "flat" models build on the overall data, and a novel stacking scheme, developed to utilize knowledge from all six datasets. When one model is built for each dataset, it turns out that whether the R-R, GSR, or merged features yield the best results is domain (dataset) dependent. When all datasets are merged into one and used to train and evaluate the models, the stacking scheme improved upon the results of the "flat" models.

Povzetek: Zaznavanje psihološkega vzburjenja iz fizioloških signalov je težka naloga, posebej če se je želimo lotiti na enoten način za več različnih domen. Vendar je potreba po inteligentnih strojih, ki so zmožni razumeti tudi čustva, vedno večja: uporabljajo se za različne probleme, od obvladovanja duševnih motenj z rešitvami mobilnega zdravstva do izboljševanja uporabniške izkušnje pri interakciji človeka z računalnikom. Odprto raziskovalno vprašanje, s katerim se ukvarja ta članek, je, katere metode strojnega učenja in kakšni vhodni podatki so primerni za zaznavanje vzburjenja. Članek opisuje več-domensko študijo zaznavanja vzburjenja na šestih različnih zbirkah podatkov. Zbirke so pretvorjene v enoten spektralno-časovni prostor intervalov R-R in galvanskega odziva kože, iz katerih izluščimo značilke in jih uporabimo kot vhod v algoritme strojnega učenja. Primerjamo modele, prilagojene posamičnim zbirkam podatkov, modele, zgrajene iz združenih podatkov vse zbirk, in inovativen ansambel modelov, ki takisto uporablja vseh šest zbirk. Izkaže se, da če zgradimo po en model za vsako zbirko podatkov, je od zbirke odvisno, ali se najbolje obnesejo značilke, izluščene iz intervalov R-R, galvanskega odziva kože ali obojega. Če zbirke podatkov združimo, pa se ansambel obnese bolje od navadnega modela.

1 Introduction

In 1897, Wundt [1] set the basis for modeling affective states by identifying the two emotional dimensions of calm-excitement and relaxation-tension. Almost a century later, in 1997, the field of affective computing [2] has been introduced, which aims for computational modeling of the affective states. Besides the maturity of the field of affective computing, modeling affective states has still remained a challenging task. Its importance is mainly reflected in the domain of human-computer interaction (HCI) and mobile health. In the HCI, it enables a natural and emotionally intelligent interaction. In the mobile health, it is used for timely detection and management of emotional and mental disorders such as depression, bipolar disorders and posttraumatic stress disorder. For example, the cost of work-related depression in Europe was estimated to €617 billion annually in 2013. The total was made up of costs resulting from absenteeism and presenteeism (€272 billion), loss of productivity (€242 billion), health care

costs of €63 billion and social welfare costs in the form of disability benefit payments (€39 billion) [3].

Affective states are complex states that results in psychological and physiological changes that influence behaving and thinking [5]. These psycho-physiological changes can be captured by a wearable device equipped with galvanic skin response (GSR – measures sweating rate), Electrocardiography (ECG – measures heart electrical activity) or blood volume pulse (BVP – measures cardiovascular dynamics) sensors. For example, the affective state of excitement usually initiates changes in heartbeat, breathing, sweating, and muscle tension, which can be captured using wearable sensors.

There are several approaches for modeling emotions, including discrete, continuous, and appraisal-driven approach. For the apprise-driven approach, context information is needed to model people's relationship to the environment that elicits their emotional response [4]. However, in computer science studies, the required context information is usually not available. In the discrete approach, the affect (emotion) is represented as discrete and distinct state, i.e., anger, fear, sadness, happiness, boredom, disgust and neutral. In the continuous approach, the emotions are represented in 2D (see Figure 1) or 3D space of activeness, valance and dominance [5]. Unlike the discrete approach, this model does not suffer from vague definitions and fuzzy boundaries, and has been widely used in affective studies [6] [7] [8]. The use of the same annotating model allows for an inter-study analysis.

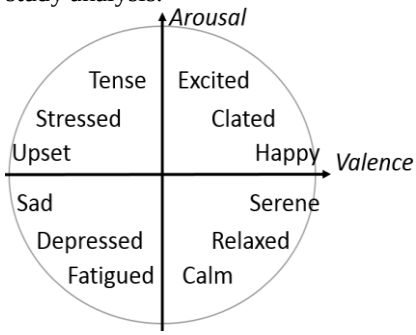

Figure 1: Circumplex model of affect. The model maps affective states in a 2D space of Arousal and Valence [5].

In this study we examine arousal recognition from GSR and heart–related physiological data, captured via: chest-worn ECG and GSR sensors, finger-worn BVP sensor, and wrist-worn GSR sensor BVP sensor. The data belongs to six publicly available datasets for affect recognition, in which there are 191 different subjects (70 females) and nearly 150 hours of arousal-labelled data. All of this introduces the problem of inter-domain learning, to which ML techniques are sensitive. To overcome this problem, we propose a preprocessing technique and a novel ML stacking scheme. The preprocessing technique translates the datasets into a common spectro-temporal space of R-R and GSR data. After the preprocessing, R-R and GSR features are extracted, which can be fed into ML algorithms to build models for arousal recognition. The novel ML stacking scheme builds dataset-specific ML models and uses a meta-learner to build general models.

The novelties of this study are:

(1) First study in affect recognition that analyzes data from six different datasets (see Section 3 Data).
(2) Methodology for translating physiological data into a common spectro-temporal space of R-R and GSR data (see Section 4.1 Pre-processing and feature extraction).
(3) Novel ML stacking scheme that generalizes from dataset-specific to general ML model for arousal recognition (see Section 4.2 Machine learning).

2 Related work

Affect recognition is an established computer-science field, but one with many remaining challenges. Many studies confirmed that affect recognition can be performed using speech analysis [10], video analysis [11], or physiological sensors in combination with ML [12]. The majority of the methods that use physiological signals use data from ECG, electroencephalogram (EEG), functional magnetic resonance imaging (fMRI), GSR, electrooculography (EOG) and/or BVP sensors.

In general, the methods based on EEG data outperform the methods based on other data [6] [7], probably due to the fact the EEG provides a more direct channel to one's mind. However, even though EEG achieves the best results, it is not applicable in normal everyday life. In contrast, affect recognition from R-R intervals or GSR data, is much more unobtrusive since this data can be extracted from ECG sensors, BVP sensors, or GSR sensors, most of which can be found in a wrist device (e.g., Empatica [13] and Microsoft Band [14]). Our methodology is tailored towards this type of data.

Regarding the typical ML approaches for affect recognition, Iacoviello et al. have combined discrete wavelet transformation, principal component analysis and support vector machine (SVM) to build a hybrid classification framework using EEG [15]. Khezri et al. used EEG combined with GSR to recognize six basic emotions via K-nearest neighbors (KNN) classifiers [16]. Mehmood and Lee used independent component analysis to extract emotional indicators from EEG, EMG, GSR, ECG and effective refractory period (ERP) [17]. Mikuckas et al. [18] presented a HCI system for emotional state recognition that uses spectro-temporal analysis only on R-R signals. More specifically, they focused on recognizing stressful states by means of the heart rate variability (HRV) analysis.

Table 1: Experimental data summary [39].

					Duration per		
Dataset	**Subjects**	**Females**	**Mean age**	**Trials**	**trial [s]**	**subject [min]**	**dataset [h]**
Ascertain	58	21	31	36	80	48.0	46.4
DEAP	32	16	26.9	40	60	40.0	21.3
Driving	10	3	35.6	1	1800	30.0	5.0
Cognitive	21	0	28	2	2400	80.0	28.0
Mahnob	30	17	26	40	80	53.3	26.7
Amigos	40	13	28	16	86	22.9	15.3
Overall	191	70	29.25	135	884.0	251.3	142.7

Regarding the more advanced ML approaches, Yin et al. [20] used an ensemble of deep classifiers for recognizing affective states using EEG, electromyography (EMG), ECG, GSR, and EOG. Using the same data, Verma et al. [19] developed an ensemble of shallow classifiers. Similarly, Kuncheva et al. [21] introduced AMBER - Advanced Multi-modal Biometric Emotion Recognition approach which uses data from EEG, EDA and HR sensor.

In contrast with the related work, which analyzes only one dataset, we perform experiments with six different datasets (domains), we analyze which ML algorithms in combination with which data type (either R-R intervals or GSR) yields best performance across all six different dataset for arousal recognition, and we propose a novel stacking method for learning from all six different domains. Finally, the work presented here is related to our previous conference paper [39]. Here we present more details regarding the data pre-processing and feature extraction, we present the novel stacking scheme and new experimental results.

3 Data

The data belongs to six publicly available datasets for affect recognition: Ascertain [6], Deap [7], Driving workload dataset [26], Cognitive load dataset [27], Mahnob [29], and Amigos [30]. Overall, nearly 150 hours of arousal-labelled data that belong to 191 subjects. Table 1 presents the data summary, which contains: number of subjects per dataset, the mean age, number of trials per subject, mean duration of each trial, duration of data per subject − in seconds, and overall duration.

Our goal was to recognize the arouse. Four datasets, Ascertain, Deap, Mahnob and Amigos, were already labelled with the subjective arousal level. One difference between these datasets was the arousal scale used for annotating. For example, the Ascertain dataset used a 7-point arousal scale, whereas the Deap dataset used a 9-point arousal scale (1 is very low, and 9 is very high, and the mean value is 5). Since the problem of arousal recognition is difficult, we decided to formulate it as a binary classification problem. From both scales, we thus split the labels in two classes using the mean value with respect the original scales. This is the same split used in the original studies. A similar step was performed for the Mahnob dataset.

Two datasets, Driving workload and Cognitive load, did not contain labels for subjective arousal level. The Driving workload dataset was labelled with subjective ratings for a workload during a driving session. For this dataset, we presume that increased workload corresponds to increased arousal. Thus, we used the workload ratings as arousal ratings. The threshold for high arousal was put on 50%. Similarly, the Cognitive load dataset was labelled for subjective stress level during stress inducing cognitive load tasks (mathematical equations). The subjective scale was from 0 to 4 (no stress, low, medium and high stress). We put the threshold for high arousal on 2 (medium stress).

4 Methods

4.1 Pre-processing and feature extraction

4.1.1 R-R data

The preprocessing is essential, since it allows merging of the six different datasets. For the heart-related data, it translates the physiological signals (ECG or BVP) to R-R intervals and performs temporal and spectral analysis. First, a peak detection algorithm is applied to detect the R-R peaks. Figure 2 presents an example for ECG signal and the detected R-R peaks. On the x-axis is the sample of the data window, on the y-axis is the output of the ECG sensor (voltage) and the detected peaks are marked with red.

Next, is temporal analysis, i.e., calculating the time distance between the detected peaks. Once the R-R intervals are detected they can be analyzed as a time series. Figure 3 is an example of an R-R time series. On the y-axis is the duration of the R-R interval, and on the x-axis is the time (in seconds) in which the R-R interval has occurred.

After the detection of R-R intervals, the R-R signal is processed. First, each R-R signal is filtered using a median filter which removes the R-R intervals that are outside of the interval [0.7*median, 1.3*median]. These parameters were determined experimentally.

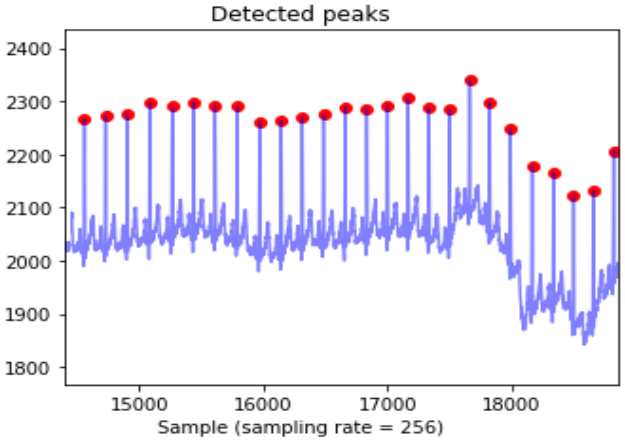

Figure 2: ECG signal and detected R-R peaks (red color). ASCERTAIN dataset t, Subject 1, Video 29 [6].

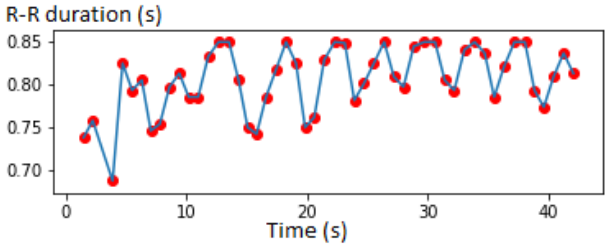

Figure 3: Example R-R signal as a time-series. ASCERTAIN dataset, Subject 1, Video 29 [6].

After the median filter, person specific winsorization is performed with the threshold parameter of 3 to remove outlier R-R intervals. From the filtered R-R signals, periodogram is calculated using the Lomb-Scargle algorithm [9]. The Lomb-Scargle algorithm allows efficient computation of a Fourier-like power spectrum estimator from unequally spaced data (as are the R-R intervals). Figure 4 presents an example Lomb-Scargle periodogram. The red color represent the low frequencies and the yellow color represents the high frequencies.

Finally, based on the related work [36], the following HRV features were calculated from the time and spectral representation of the R-R signals: the mean heart rate (meanHR), the mean of the R-R intervals (meanRR), the standard deviation of the R-R intervals (sdnn), the standard deviation of the differences between adjacent R-R intervals (sdsd), the square root of the mean of the squares of the successive differences between adjacent R-R intervals (rmssd), the percentage of the differences between adjacent R-R intervals that are greater than 20 ms, the percentage of the differences between adjacent R-R intervals that are greater than 50 ms, Poincaré plot indicies (SD1 and SD2), total spectral power of all R-R samples in power between 0.003 and 0.04 Hz (lf - low frequencies), between 0.15 and 0.4 Hz (hf − high frequencies), and the ratio of low to high frequency power.

4.1.2 GSR data

To merge the GSR data from the six datasets, several problems were addressed. Each dataset is recorded with

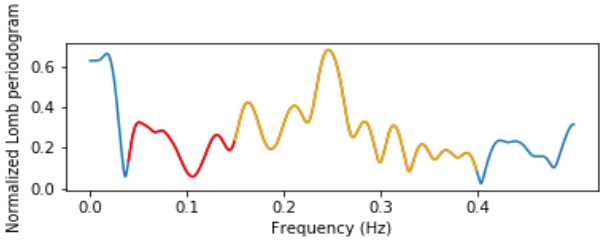

Figure 4: Normalized Lomb-Scargle periodogram calculated from R-R signal. ASCERTAIN dataset, Subject 1, Video 29 [6].

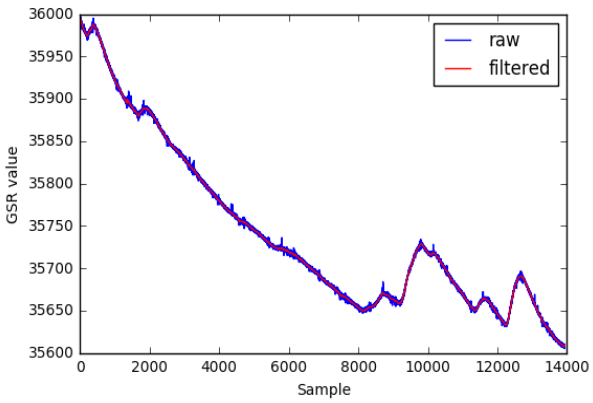

Figure 5: Filtered GSR signal. ASCERTAIN dataset, person 1, Clip 1 [6].

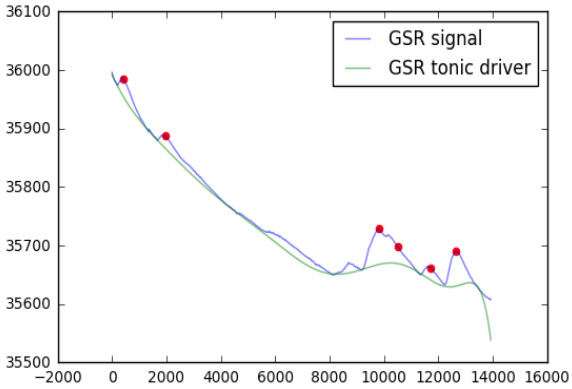

Figure 6: GSR signal decomposition (green – tonic driver, slow acting component; red – GSR responses, fast acting component). ASCERTAIN dataset, person 1, Clip 1 [6].

different GSR hardware, thus the data can be presented in different units and different scales. To address this problem, each GSR signal was converted to μS (micro Siemens). Next, the GSR signal was filtered using a lowpass filter with a cut-off frequency of 1 Hz. Figure 5 presents an example filtered GSR signal. To address the inter-participant variability of the signal, person-specific min-max normalization was performed, i.e., each signal was scaled to [0, 1] using person specific winsorized minimum and maximum values. The winsorization parameter was set to 3.

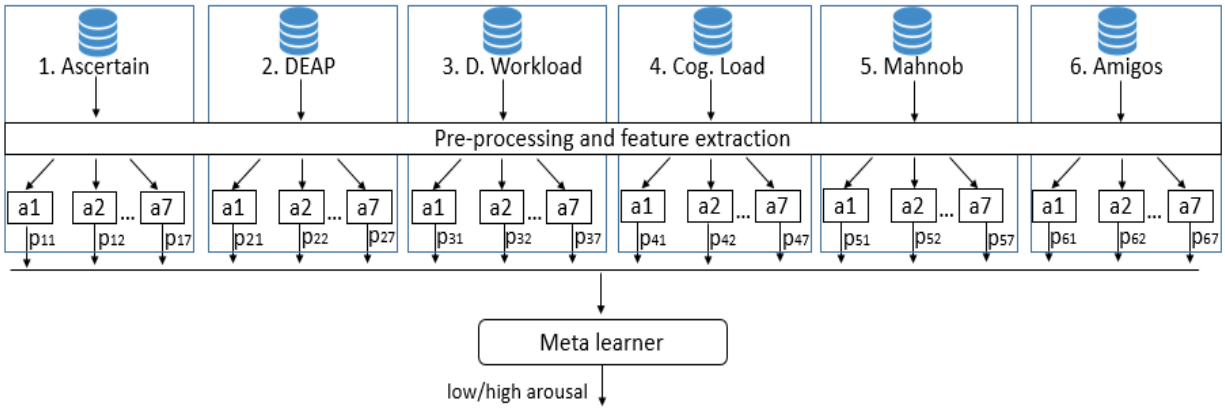

Figure 7: The novel stacking scheme for training a meta-learner that utilizes knowledge from all six datasets.

Finally, the fast acting component (GSR responses) and the slow acting component (tonic component) were determined in the signal using the "peakutils.baseline" function from the Python's PeakUtils library. The function is used with the default parameters. It iteratively performs a polynomial fitting in the data to detect its baseline. For example, in Figure 6, the GSR responses are marked with red and the tonic component (baseline) is marked with green. Based on the related work [30], the preprocessed GSR signal was used to calculate GSR features: mean, standard deviation, 1st and 3rd quartile (25th and 75th percentile), quartile deviation, derivative of the signal, sum of the signal, number of responses in the signal, rate of responses in the signal, sum of the responses, sum of positive derivative, proportion of positive derivative, derivative of the tonic component of the signal, difference between the tonic component and the overall signal.

4.2 Machine learning

4.2.1 Flat machine learning

After the feature extraction, the data is in a format which can be input for typical ML algorithms. Models were built using seven different ML algorithms: Random Forest, Support Vector Machine, Gradient Boosting Classifier, AdaBoost Classifier (with a Decision Tree as a base classifier), KNN Classifier, Gaussian Naive Bayes and Decision Tree Classifier. The algorithms were used as implemented in the Scikitlearn, the Python ML library [37]. For each algorithm, a randomized search on hyper parameters was performed on the training data using 2-fold cross-validation.

4.2.2 Stacking

The novel stacking scheme, depicted in Figure 7, was designed to train a meta-learner which would utilize the knowledge from all six datasets. In the example scenario, we used the 7 ML algorithms mentioned in the previous section. Thus, there are 42 base models (6 datasets x 7 ML algorithms). The outputs of the base models, which are probabilities for the class "high arousal", are used as input to a meta-learner. The meta-learner can be any ML algorithm previously mentioned. We experimentally chose Random Forest to be our meta-learner. The meta-learner is trained using a 10 fold-cross validation on the training data. That is, the base learners are trained on 90% of the data, then predictions are provided on the rest 10% of the data, and this procedure is repeated ten times. Finally, the meta-learner is trained on the cross-validated predictions of the base learners. In the test phase, the test instances are provided as input to all of the 42 base models, their output is summed up in a 42 dimensional vector (in Figure 7 marked as p_{11}, p_{12},…,p_{67} – six datasets and seven base models) as input to the meta-learner, which provides the final prediction for the test instance.

5 Experimental results

Two types of experiments were performed: dataset specific experiments, and experiments with merged datasets. The dataset-specific experiments were used to identify the ML algorithm and the input that would yield the best performance per dataset.

The experiments on the merged datasets were used to build general, dataset-independent ML models. This evaluation simulates a scenario where the source (dataset) is unknown, i.e., we do not know whether the subject is watching an affective video (e.g., the DEAP dataset), is driving a car (e.g., the Driving workload dataset) or he/she is working on a cognitive demanding task (e.g., the cognitive load dataset).

The evaluation was performed using trial-specific 10-fold cross-validation, i.e., the data segments that belong to one trial (e.g., one affective stimuli), can either belong only to the training set or only to the test set, thus there was no overlapping between the training and test data.

5.1 Dataset specific

The results for the dataset-specific experiments are presented in Table 2. The first column represents the ML algorithm, the second column represents the features used as input to the algorithm (R-R, GSR or Merged - M) and the rest of the columns represent the dataset which is used for training and evaluation using the trial-dependent 10-fold cross-validation. We report the mean accuracy ± the standard evaluation for the 10 folds. For each dataset, the best performing model(s) is (are) marked with green.

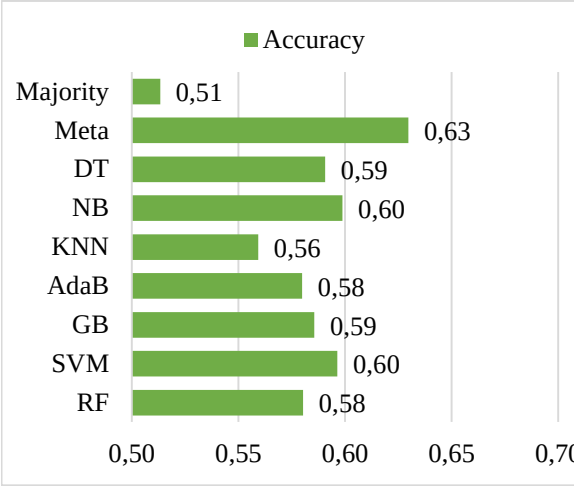

Figure 8: Accuracy of the meta-learner and the "flat" approaches for the merged-datasets experiments.

For example, on the Ascertain and the Driving workload dataset, the best performing algorithm is the SVM, on the Deap dataset, the best performing algorithm is the RF, on the Cognitive Load and the Mahnob datasets, the best performing is the NB, and on the Amigos dataset, the best performing is the AdaBoost algorithm.

When we compare which input (R-R features, GSR features or Merged-M) provide better accuracy, on two datasets, the Asceratin and the Driving workload, the results are the same, on the Deap dataset, the R-R features provide better results, on the Cognitive Load dataset, the highest accuracy is achieved both for the GSR and the Merged features, on the Mahnob dataset, the GSR features provide best accuracy and on the Amigos dataset, the Merged features.

Regarding the majority class, the biggest accuracy improvement was achieved for the Cognitive load dataset, which is an improvement of 9 percentage points. For the two datasets, the Deap and the Amigos, the improvement was 2-3 percentage points, and for the three datasets, the Ascertain, the Driving workload and the Mahnob, the best performing models were as good as the majority classifier.

5.2 Merged datasets

In the dataset-specific experiments, none of the algorithms yielded best performance (compared to the rest of the algorithms) over all datasets, thus there was no experimental hint about which algorithm would be able to generalize over all datasets. For that reason, we came up with the stacking approach, where a meta-learner learns how to combine the output of all of the algorithms trained on the different datasets. The details are presented in section 4.2. Stacking. The input to the algorithms was the merged feature set, i.e., R-R and GSR features.

We compared the meta-learning approach to a simple approach where the "flat" ML algorithms are trained on all datasets merged. The evaluation is performed using the same trial-specific 10-fold cross-validation. The results are presented in Figure 8. It can be seen that all of the "flat" algorithms achieved an accuracy bellow or equal to 60%. The meta-learning approach slightly improved the results by achieving an

Table 2: Dataset-specific experimental results. Mean accuracy ± stdDev for trial-specific 10-fold cross validation. The best performing models per dataset are marked with green [39].

Algorithm	Features	Dataset					
		Ascertain	Deap	D. Workload	Cog. Load	Mahnob	Amigos
RF	R-R	0.655 ± 0.07	0.556 ± 0.03	0.785 ± 0.24	0.739 ± 0.13	0.580 ± 0.11	0.536 ± 0.06
	GSR	0.638 ± 0.06	0.503 ± 0.04	0.780 ± 0.24	0.763 ± 0.12	0.611 ± 0.07	0.473 ± 0.11
	M	0.653 ± 0.05	0.540 ± 0.04	0.785 ± 0.25	0.755 ± 0.13	0.611 ± 0.10	0.559 ± 0.10
SVM	R-R	0.664 ± 0.07	0.536 ± 0.05	0.795 ± 0.26	0.717 ± 0.21	0.623 ± 0.15	0.521 ± 0.24
	GSR	0.664 ± 0.07	0.525 ± 0.05	0.795 ± 0.26	0.712 ± 0.20	0.588 ± 0.10	0.470 ± 0.12
	M	0.664 ± 0.07	0.513 ± 0.03	0.795 ± 0.26	0.691 ± 0.18	0.623 ± 0.15	0.506 ± 0.13
GB	R-R	0.649 ± 0.07	0.554 ± 0.03	0.785 ± 0.20	0.736 ± 0.15	0.578 ± 0.11	0.543 ± 0.06
	GSR	0.642 ± 0.05	0.500 ± 0.04	0.800 ± 0.21	0.743 ± 0.12	0.609 ± 0.08	0.527 ± 0.09
	M	0.644 ± 0.05	0.533 ± 0.03	0.755 ± 0.23	0.761 ± 0.15	0.609 ± 0.11	0.542 ± 0.09
AdaB	R-R	0.658 ± 0.06	0.532 ± 0.02	0.750 ± 0.23	0.718 ± 0.13	0.580 ± 0.09	0.531 ± 0.07
	GSR	0.633 ± 0.05	0.485 ± 0.03	0.750 ± 0.22	0.740 ± 0.13	0.589 ± 0.08	0.514 ± 0.09
	M	0.623 ± 0.05	0.526 ± 0.03	0.755 ± 0.22	0.766 ± 0.16	0.610 ± 0.08	0.560 ± 0.08
KNN	R-R	0.625 ± 0.05	0.509 ± 0.02	0.710 ± 0.19	0.715 ± 0.13	0.582 ± 0.07	0.509 ± 0.05
	GSR	0.590 ± 0.06	0.496 ± 0.04	0.795 ± 0.26	0.772 ± 0.09	0.605 ± 0.06	0.533 ± 0.08
	M	0.600 ± 0.05	0.490 ± 0.02	0.750 ± 0.23	0.770 ± 0.13	0.601 ± 0.09	0.533 ± 0.06
NB	R-R	0.654 ± 0.07	0.537 ± 0.04	0.735 ± 0.15	0.748 ± 0.15	0.574 ± 0.06	0.485 ± 0.09
	GSR	0.602 ± 0.04	0.537 ± 0.05	0.540 ± 0.22	0.803 ± 0.09	0.624 ± 0.07	0.454 ± 0.10
	M	0.591 ± 0.04	0.535 ± 0.06	0.665 ± 0.17	0.804 ± 0.12	0.592 ± 0.06	0.486 ± 0.09
DT	R-R	0.664 ± 0.07	0.519 ± 0.05	0.685 ± 0.17	0.736 ± 0.15	0.597 ± 0.09	0.505 ± 0.06
	GSR	0.640 ± 0.05	0.542 ± 0.05	0.765 ± 0.22	0.734 ± 0.08	0.583 ± 0.09	0.483 ± 0.11
	M	0.650 ± 0.05	0.524 ± 0.04	0.615 ± 0.22	0.704 ± 0.09	0.581 ± 0.13	0.551 ± 0.09
Majority		0.664	0.536	0.795	0.717	0.623	0.521

accuracy of 63%.

6 Conclusion and discussion

We presented an inter-domain study for arousal recognition on six different datasets, recorded with twelve different hardware sensors. We experimented with dataset-specific models, general models built on the overall (merged) data and general models build using the novel stacking scheme. For the dataset-specific models, we compared the results of seven different ML algorithms, using three different feature inputs (R-R, GSR or Merged – M features). For the models built on the overall (merged) data, we compared the results of the novel stacking scheme and "flat" ML models. The results on the dataset-specific setup showed that, out of the seven ML algorithms tested, none yields the best performance on all datasets. In addition to that, a clear conclusion cannot be made whether the R-R, GSR or the Merged features yield the best results – this is domain (dataset) dependent.

On the merged-datasets experiments, the novel stacking scheme slightly outperformed the "flat" models. This was expected since the stacking scheme utilizes seven different ML models built on the six different datasets, thus 42 different models (views).

However, the experimental results show that there is room for improvement regarding the accuracy achieved in both types of experiments. In the future, we plan to investigate more advanced techniques such as deep neural networks and transfer learning, which might be able to learn more accurate models that will be able to generalize across different domains. Finally, once we find the best performing scenario, we will generalize the method for arousal recognition to a method for valence recognition and method for discrete emotion recognition.

7 References

[1] W. Wundt. Outlines of psychology (C. H. Judd, Trans.). Oxford, UK: Engelman, 1897.

[2] R. Picard. Affective Computing. Cambridge, MA: MIT Press, 1997.

[3] Depression cost: http://ec.europa.eu/health/sites/health/files/mental_health/docs/matrix_economic_analysis_mh_promotion_en.pdf, [Accessed 27.03.2017].

[4] S. Marsella, J. Gratch. Computationally modeling human emotion. Commun. ACM 57, 12 (November 2014), pp. 56-67. 2014.

[5] J. A. Russell. A circumplex model of affect. Journal of Personality and Social Psychology, 1980.

[6] R. Subramanian, J. Wache, M. Abadi, R. Vieriu, S. Winkler, N Sebe. ASCERTAIN: Emotion and Personality Recognition using Commercial Sensors. IEEE Transactions on Affective Computing. 2016.

[7] S. Koelstra, C. Muehl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras. DEAP: A Database for Emotion Analysis using Physiological Signals (PDF). IEEE Transaction on Affective Computing, 2012.

[8] M.K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras. Nicu Sebe. DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses. IEEE Transactions on Affective Computing, 2015.

[9] N.R. Lomb. Least-squares frequency analysis of unequally spaced data. Astrophysics and Space Science, vol 39, pp. 447-462, 1976

[10] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang.A survey of affect recognition methods: audio, visual, and spontaneous expressions. Transactions on Pattern Analysis and Machine Intelligence, 2009.

[11] I. Abdic, L. Fridman, D. McDuff, E. Marchi, B. Reimer, Schuller, B. Driver Frustration Detection From Audio and Video. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'16), 2016.

[12] S. Jerritta, M. Murugappan, R, Nagarajan, K. Khairunizam. Physiological Signals Based Human Emotion Recognition: A Review. International Colloquium on Signal Processing and its Applications. 2011.

[13] M. Garbarino, M. Lai, D. Bender, R. W. Picard, S. Tognett. Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. 4th International Conference on Wireless Mobile Communication and Healthcare, pp. 3-6, 2014.

[14] Microsoft band. https://www.microsoft.com/microsoft-band/en-us

[15] D. Iacovielloa, A. Petraccab, M. Spezialettib, G. Placidib. A real-time classification algorithm for EEG-based BCI driven by self-induced emotions. Computer Methods and Programs in Biomedicine, 2015.

[16] M. Khezria, M.Firoozabadib, A. R. Sharafata. Reliable emotion recognition system based on dynamic adaptive fusion of forehead biopotentials and physiological signals.

[17] R. M. Mehmooda, H. J. Leea. A novel feature extraction method based on late positive potential for emotion recognition in human brain signal patterns. Computers & Electrical Engineering, 2016.

[18] A. Mikuckas, I. Mikuckiene, A. Venckauskas, E. Kazanavicius2, R. Lukas2, I. Plauska. Emotion Recognition in Human Computer Interaction Systems. Elektronika Ir Elektrotechnika, Reserarch Journal, Kaunas University of Technology, 2014.

[19] G. K. Verma, U. S. Tiwary. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. NeuroImage, 2014.

[20] Z. Yin, M. Zhao, Y. Wang, J. Yang, J Zhang. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. Computer Methods and Programs in Biomedicine, pp. 93-110, 2017.

[21] L. I. Kuncheva, T. Christy, I. Pierce, Sa'ad P. Mansoor. Multi-modal Biometric Emotion Recognition Using Classifier Ensembles.

Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2011.
[22] Wei Liu, Wei-Long Zheng, Bao-Liang Lu. Multimodal Emotion Recognition Using Multimodal Deep Learning. Online. Available at: https://arxiv.org/abs/1602.08225, 2016.
[23] W-L. Zheng, B-L Lu. A multimodal approach to estimating vigilance using EEG and forehead EOG. Journal of Neural Engineering, 2017.
[24] Z. Yin, M. Zhao, Y. Wang, J. Yang, J. Zhang. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. Comput Methods Programs Biomed. 2017.
[25] K.Weiss, T. M. Khoshgoftaar, D. Wang. A survey of transfer learning. Journal of Big Data, 2016.
[26] S. Schneegass, B. Pfleging, N. Broy, A. Schmidt, Frederik Heinrich. A Data Set of Real World Driving to Assess Driver Workload. 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 2013.
[27] M. Gjoreski, M. Luštrek, M. Gams, H. Gjoreski. Monitoring stress with a wrist device using context. Journal of Biomedical Informatics, 2017, in press.
[28] M. Gjoreski, H. Gjoreski, M. Luštrek, M. Gams. Continuous stress detection using a wrist device: in laboratory and real life. ACM Conf. on Ubiquitous Computing, Workshop on mentalhealth, pp. 1185-1193, 2016.
[29] M. Soleymani, T.Pun. A Multimodal Database for Affect Recognition and Implicit Tagging, IEEE Transactions On Affective Computing, 2012.
[30] J. A. Miranda-Correa, M. Khomami Abadi, N. Sebe, I. Patras. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. Transactions On Affective Computing, 2017.
[31] L. H. Negri. Peak detection algorithm. Python Implementation. Online. Available at: http://pythonhosted.org/PeakUtils/.
[32] M. Wu, PhD thesis. Michigan State University; 2006. Trimmed and Winsorized Eestimators.
[33] J.D. Scargle. Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. The Astrophysical Journal, vol 263, pp. 835-853, 1982.
[34] D. P. Kingma, J. Ba. Adam: A Method for Stochastic Optimization, http://arxiv.org/abs/1412.6980, 2014.
[35] Tensorflow. Online. Available at: https://www.tensorflow.org/
[36] R. Castaldoa, P. Melillob, U. Bracalec, M. Casertaa,c, M. Triassic, L. Pecchiaa. Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis. Biomedical Signal Processing and Control. 2015.
[37] Scikit-learn, Python machine-learning library http://scikit-learn.org/dev/_downloads/scikit-learn-docs.pdf
[38] L.J.P, van der Maaten., G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research. 9: 2579–2605, 2008.
[39] M. Gjoreski, B. Mitrevski, Mitja Luštrek, Matjaž Gams. R-R vs GSR – An inter-domain study for arousal recognition, Multiconference Information Society, Ljubljana, 2017.
[40] Python library for signal analysis: http://pythonhosted.org/PeakUtils/