Proceedings of Machine Learning Research 86:52–65, 2018 Workshop on Artificial Intelligence in Affective Computing

Deep Ensembles for Inter-Domain Arousal Recognition

Martin Gjoreski MARTIN.GJORESKI@IJS.SI
Department of Intelligent Systems, Jozef Stefan Institute, Ljubljana, Slovenia

Hristijan Gjoreski
Faculty of Electrical Engineering, University Ss. Cyril and Methodius, Skopje, Macedonia

Mitja Luštrek
Matjaž Gams
Department of Intelligent Systems, Jozef Stefan Institute, Ljubljana, Slovenia

Abstract

The computer science field Affective Computing, which studies and develops emotional intelligent systems, has been active for almost two decades now with limited results. Arousal as the dimension that represents the intensity of the emotions, represents similar recognition problems. This is the first study that analyzes six publicly available datasets for arousal recognition from physiological signals and proposes a method capable of combining them. The novel method, an inter-domain Deep Neural Network (DNN) ensemble, is compared to classical machine learning (ML). For both methods, the raw data from Galvanic Skin Response (GSR), Electrocardiography ECG, and Blood Volume Pulse (BVP) sensors is processed and transformed into a common spectro-temporal space of R-R intervals and GSR data. For the classical ML algorithms, features are extracted, and for the DNN algorithms, two different approaches were taken: a fully connected DNN trained with the same features as the classical ML algorithms (DNN-Features) and a Convolutional Neural Network (CNN) trained with the temporal representation of the GSR signal (CNN-GSR). Finally, a fully connected DNN meta learner is trained to utilize the knowledge from the two different DNNs and to tune the DNN models for the target dataset. The experimental results showed that the novel DNN ensemble method outperforms the classical ML methods and the non-ensemble DNN methods. Additionally, the CNN-GSR model learned that the peaks of the GSR signal contain the most information regarding the arousal, thus the network developed filters to emphasize those parts.

Keywords: Affective computing, affect recognition, deep learning, convolutional networks, ensembles

1. Introduction

Emotions are paramount in the human communication. They serve as a medium to enrich the communication, to express preferences, to communicate subjective cues, and even to manipulate others. The book "Affective Computing" from Picard (1997) is considered as the birth of the scientific field that studies and develops emotion aware computer systems. Two decades afterwards with Affective Computing as a well-established research field, modeling emotional states still remains a challenging task.

With the advancement of the technology and the penetration of the information systems into our everyday life, the need for emotion-aware systems is becoming increasingly more

evident. For example, in the domain of human-computer interaction (HCI), an emotion-aware system would enable a more natural interaction and better user experience. In the healthcare domain, a system for monitoring emotion can contribute to the timely detection and treatment of emotional and mental disorders such as depression, bipolar disorders and posttraumatic stress disorder (PTSD). From an economical point of view, emotion monitoring system may help to decrease the cost of work-related depression. The cost in 2013 in Europe was estimated to 617 billion annually [1].

One popular approach for modeling emotions in psychology, is to represent the emotions in a 2D or 3D space of arousal, valance and dominance Russell (1980). This approach takes into account the vague definitions and fuzzy boundaries of the emotional states, and has been widely used in affective studies for annotating data Koelstra et al. (2012); Subramanian et al. (2017). The use of the same psychological approaches for annotating data across multiple computer-science studies related to emotion recognition allows for an inter-domain analysis. More specifically, we analyze six publicly available datasets for affect recognition with 142 hours of arousal-labelled data that belongs to 191 subjects (70 females and 121 males). We focus on arousal recognition from physiological data captured via chest-worn Electrocardiography (ECG) sensors, finger-worn wrist-worn blood volume pulse (BVP) sensors, and chest-worn and wrist-worn Galvanic Skin Response (GSR) sensor.

Even though the same type of data were recorded in the six datasets, different sensors were used for each of them, resulting in variations in the data collected. For example, the GSR sensor is different for each dataset. To overcome this problem, we exploit a four-step solution. First, the data from the ECG and BVP sensors is preprocessed and transformed into a common spectro-temporal space of R-R intervals and Lomb-Scargle periodogram Lomb (1976), regardless of the sensor. Second, the data from the GSR sensors is normalized and transformed into the same unit (micro Siemens). Third, we propose a meta-learner which is specifically tuned for each dataset. Finally, to exploit the knowledge from all six datasets we used DNN techniques for learning arousal-recognition models, since DNN approaches are known to scale much better than classical flat ML algorithms on large datasets. More specifically, we trained DNN models in the feature space of R-R and GSR features, and CNN models directly on the raw GSR data to capture some additional information that may have diminished in the feature extraction phase.

Highlights of the study: (*i*) Preprocessing methods for translating different datasets into a common spectro-temporal space, paving the way for further inter-domain studies exploiting the data accumulated by the ubiquitous computing community; (*ii*) A novel DNN ensemble method for arousal recognition that benefits from large amounts of data even when the data are heterogeneous (i.e., 191 different subjects, twelve different sensors, six different datasets and three different placements), which outperforms flat classical ML and meta-ML approaches; (*iii*) Comparison of classical ML and deep ML approaches for arousal recognition on six different datasets; (*iv*) New insights from the CNN-GSR model. The CNN-GSR model discovered that the peaks of the GSR signal are the parts that contain the most information regarding the arousal, thus the network developed filters to stress those parts of the signals.

1. http://ec.europa.eu/health//sites/health/files/mental_health/docs/matrix_economic_analysis_mh_promotion_en.pdf

2. Related Work

Affect recognition is an established computer-science field, but one with many challenges remaining. There has been many studies confirming that affect recognition can be performed using speech analysis Trigeorgis et al. (2016), video analysis Subramanian et al. (2017), or physiological sensors in combination with ML. The majority of the methods that use physiological signals use data from ECG, electroencephalogram (EEG), functional magnetic resonance imaging (fMRI), galvanic skin response (GSR), electrooculography (EOG) and/or BVP sensors. In general, the methods based on EEG data outperform the methods based on other data Subramanian et al. (2017), probably beacuse 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 may be much more unobtrusive since R-R intervals can be extracted from ECG sensors or BVP sensors, including sensors in a wrist device (e.g., Empatica Garbarino et al. (2014) and Microsoft Band [2]). Regarding the typical ML approaches for affect recognition, Iacoviello et al. (2015) have combined discrete wavelet transformation, principal component analysis and support vector machine (SVM) to build a hybrid classification framework using EEG. Khezri et al. (2015) used EEG combined with GSR to recognize six basic emotions via K-nearest neighbors (KNN) classifiers. Verma and Tiwary (2014) developed an ensemble approach using EEG, electromyography (EMG), ECG, GSR, and EOG. Mehmood and Lee (2016) used independent component analysis to extract emotional indicators from EEG, EMG, GSR and ECG .

Recently, the use of deep learning for affect recognition has become popular, too. Liu et al. (2016) presented a deep learning approach for emotion recognition using EEG data and eye blink data. Similarly, Bashivan et al. (2015) presented an approach for learning representations from EEG signal with deep recurrent-convolutional neural networks. Yin et al. (2017) presented an approach for the recognition of emotions using multimodal physiological signals and an ensemble deep learning model using EEG, EMG, ECG, GSR, EOG, BVP, respiration rate and skin temperature. In contrast to the EEG-based methods for affect recognition, Martinez et al. (2013) presented a DNN method for affect recognition from GSR and BVP data.

The related work shows that – similarly to many other fields – deep learning has the potential to outperform classical ML in affect recognition. However, the work done so far could not take full advantage of deep learning because training a DNN model requires a large amount of data, and yet, most of the studies in the related work analyze small datasets (10 - 50 subjects). The size of these datasets is far from the size of the datasets used in other fields (e.g. ImageNet contains 1.2 million images). To overcome this challenge, we explore inter-dataset meta DNN approach.

3. DATA

The data is presented Table 1. It consists of six publicly available datasets for affect recognition: Ascertain - Subramanian et al. (2017), Deap - Koelstra et al. (2012), Driving workload dataset - Schneegass et al. (2013), Cognitive load dataset - Gjoreski et al. (2017)

2. https://www.microsoft.com/microsoft-band/en-us

Table 1: Description of the datasets in the study

dataset	# subjects	age	# trials	trial[s]	subject[min]	dataset[h]
Ascertain	58	31	36	80	48	46.4
DEAP	32	26.9	40	60	40	21.3
Driving	10	35.6	1	1800	30	5
Cognitive	21	28	2	2400	80	28
Mahnob	30	26	40	80	53.3	26.7
Amigos	40	28	16	86	22.9	15.3
Overall	191	29.25	135	884	251.3	142.7

, Mahnob - Soleymani et al. (2012), and Amigos - Abdon Miranda-Correa et al. (2017). Overall, 142 hours of arousal-labelled data is presented belonging to 191 subjects. The rest of the columns present: the number of subjects per dataset, the mean age, the number of trials per subject, the mean duration of each trial, the duration of the data per subject, and the overall duration.

Our goal was to recognize the arousal. Four datasets (Ascertain, Deap, Mahnob and Amigos) were already labeled with the subjective arousal level. In these studies (datasets), the subjects were watching affective videos with an average duration of 60-80 seconds. After each video, the subjects took a rest and filled questionnaires that include subjective arousal ratings. The ratings are used as labels for the physiological data. However, these datasets use different arousal scale 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 to the original scales. This is the same split used in the original studies. A similar step was performed for the Mahnob and the Amigos dataset.

In the Driving workload dataset, the subjects had a 30-minute driving session which included highway and city driving. Each subject rated their own driving session post-driving by watching a video recording of their driving session. For this dataset, we presume that increased workload corresponds to increased arousal. Thus, we used the workload ratings as arousal ratings. Same as with the previous datasets, the mean workload value was used to split the labels in two classes, high and low arousal.

The Cognitive load dataset was labelled for subjective stress level during stress inducing cognitive load tasks. A series of randomly generated equations were presented to subjects, who provide answers verbally. The time given per equation was dynamically changing. Each session consisted of three series of equations with increasing difficulty: easy, medium and hard. After each session, the subjects took a short rest and filled questionnaires that include subjective stress rating. The subjective scale was from 0 to 3 (no stress, low, medium and high stress). Same as with the previous datasets, the mean value was used to split the labels in two classes, high and low arousal. Figure 1 depicts the distribution of each dataset after the binary split.

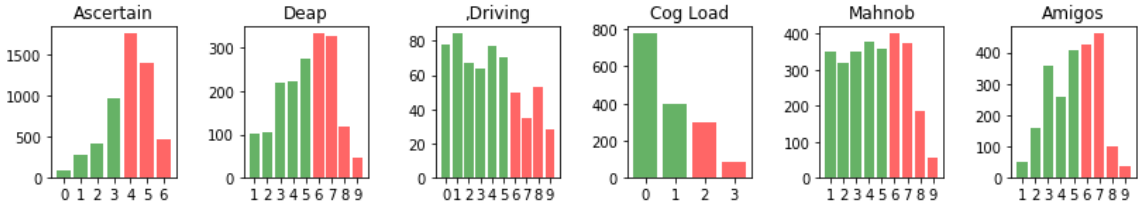

Figure 1: Histograms for each dataset in the study. Green – low arousal, red – high arousal.

4. METHODS

The novel DNN ensemble method is depicted in Figure 2. The method utilizes six different AC datasets which contain data from GSR, and ECG or BVP. First, the datasets are processed and transformed into a common spectro-temporal space of R-R intervals (extracted from the ECG and BVP data) and GSR data. After the preprocessing, two different approaches were utilized:

1. DNN-Features approach, which includes feature extraction and application of DNN. The feature extraction process extracts two types of features: (i) from the R-R intervals using heartrate variability (HRV) analysis, (ii) from the GSR signals using peak analysis and decomposition of the GSR signal into a slow-acting and a fast-acting component. The extracted features are then fed into a fully connected DNN (DNN-Features) to build models for arousal recognition.

2. CNN-GSR approach, which uses the processed GSR signals as input into a CNN. The CNN-GSR contains 2 convolutional layers, which serve as a feature extractor for two fully connected layers placed at the end of the CNN-GSR network.

Finally, a fully connected DNN meta learner is trained to utilize the knowledge from the two different DNNs (DNN-Features and CNN-GSR). The technical details for each step are explained in the following subsections.

4.1. Preprocessing and Feature extraction

4.1.1. R-R data

The preprocessing is the first essential step. It addresses the variations in the sensor data across the different datasets and allows us to merge the six datasets. For the heart-related data, it transforms the physiological signals (ECG or BVP) to R-R intervals and performs temporal and spectral analysis. The first preprocessing step is the removing of the trend of the ECG and the BVP signals. Trend is the change of the mean of the signal over time. The left graph in Figure 3 presents a BVP signal with changing trend over time, and the middle graph in Figure 3 presents the same BVP signal after the detrending.

Next, Negri's peak detection algorithm [3] is applied to detect the R-R peaks. The right graph in Figure 3 presents the BVP signal with the detected R-R peaks. After the R-

3. http://pythonhosted.org/PeakUtils/

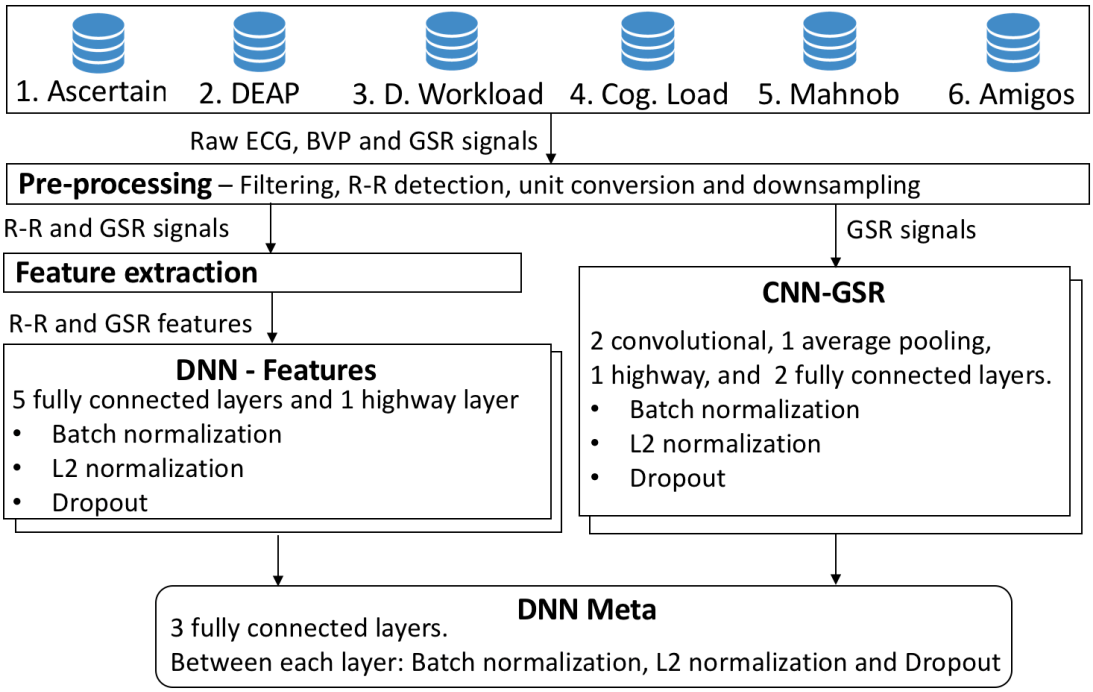

Figure 2: The proposed DNN ensemble method for arousal recognition.

R detection, each R-R signal filtered by removing the R-R intervals that are outside of the interval [0.7*median, 1.3*median]. Next, a person-specific winsorization is performed by removing the outliers outside the range [3rd, 97th] percentile. From the filtered R-R signals, a spectral representation is calculated using the Lomb-Scargle algorithm. The detailed information about the algorithms and their parameters can be found in our previous publication Gjoreski et al. (2018). Finally, based on the related work Castaldo et al. (2015), the following HRV features were calculated from the time and spectral representation of the R-R signals: the mean heart rate, the mean of the R-R intervals, the standard deviation of the R-R intervals, the standard deviation of the differences between adjacent R-R intervals, the square root of the mean of the squares of the successive differences between adjacent R-R intervals, 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, the total spectral power of all R-R samples between 0.003 and 0.04 Hz (lf - low frequencies) and 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 different GSR hardware, thus the data is presented in different units and different scales. To address this problem, each GSR signal was converted to S. Next, the GSR signal was filtered using a lowpass filter with a cut-off frequency of 1 Hz.

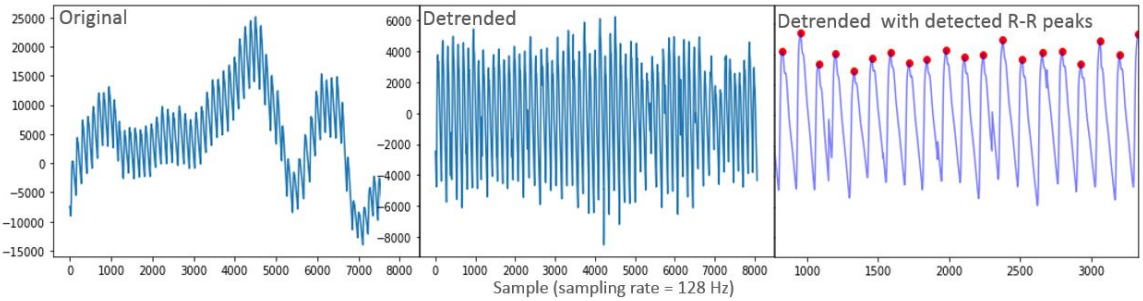

Figure 3: (left) Raw BVP signal over time. (middle) BVP signal after detrending. (right) BVP signal with RR-peaks detected.

To address the inter-participant variability each signal was scaled to [0, 1] using person-specific winsorized minimum and maximum values. Finally, the fast-acting component (GSR responses) and the slow acting component (tonic component) were extracted from the filtered GSR signals. Based on the related work Soleymani et al. (2012), 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, responses per minute 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. One additional problem that we addressed to train the CNN-DNN network is the fact that each GSR sensor had its own sampling frequency (some had 128 Hz, some 256 Hz) and trials are of varying lengths. For example, one trial in which the subjects watch a scary video may last 60 seconds and another may last 80 seconds. Thus, the recorded GSR data is of varying lengths also. For this reason, we used an undersampling algorithm which takes into account the length and the frequency of the data and transforms it into a vector of 100 samples. Thus, the GSR data of each trial (instance) is represented by a vector of 100 samples. These vectors are used as input to the CNN-DNN.

4.2. Deep ML

4.2.1. DNN-Features

The input to the DNN-Features models was the same as the input for the classical flat ML models, i.e., the extracted features from the R-R and the GSR signals. For training the models we used five fully connected DNN layers and one highway layer placed after the second fully connected layer. Each layer employed rectified linear units (ReLUs). To avoid overfitting, L2 regularization and dropout methods were used. The dropout probability was set to 0.25 and the L2 regularization rate was set to 10^{-3}. After each layer, batch normalization was used to avoid internal covariance shift Ioffe and Szegedy (2015). The batch normalization step normalizes the activations of the previous layer at each batch, i.e., applies a transformation that maintains the mean activation close to 0 and the activation

standard deviation close to 1. The training was performed by backpropagating the gradients through all layers. The parameters were optimized by minimizing the crossentropy loss function using the ADAM optimizer Kingma and Ba (2014). Learning rate of 10^{-3} and a decay rate of 10^{-2} was used. The batch size was set to 256 and the maximum number of training epochs was set to 256. The output of the model is obtained from the final layer with a softmax activation function yielding a class probability distribution.

4.2.2. CNN-GSR

We used two convolutional layers separated by one average pooling layer. Each convolutional layer contained 32 kernels with kernel-size set to 5 and stride-size set to 1. For the pooling layer, the pool size was set to 3 and the stride size was set to 1. The output of the second convolutional layer was input to a highway layer. After the highway layer, a batch normalization was performed, and the normalized outputs were input to a fully-connected layer with size of 64 kernels. Each layer employed ReLUs. To avoid overfitting, L2 regularization and dropout methods were used for the non-convolutional layers. The dropout probability was set to 0.25 and the L2 regularization rate was set to 10^{-3}. Gradient backpropagation and ADAM optimizer with a learning rate of 10^{-4} and a decay rate of 10^{-4} were used for training the models. The batch size was set to 256 and the maximum number of training epochs was set to 256. The output of the model is obtained from the final layer with a softmax activation function yielding a class probability distribution.

4.2.3. DNN-Meta

The outputs of the DNN-Features and CNN-GSR models were used ‘as input to the DNN-Meta model. We used a DNN with two hidden layers. Each layer employed ReLUs with L2 regularization rate of 10^{-3}. Between each layer, batch normalization was used. Gradient backpropagation and ADAM optimizer with a learning rate of 10^{-3}. The batch size was set to 128 and the maximum number of training epochs was set to 100. The output of the model is obtained from the final layer with a softmax activation function. The output of this model is presented as the final prediction of the algorithm. All neural networks were implemented using Tensorflow [4] and Keras [5].

5. Experiments

We compared our novel DNN-based method to classical ML methods. That is, once the features were extracted we applied classical ML methods in order to create models for arousal recognition. Models were built using seven different ML algorithms: Random Forest, Support Vector Machine, Gradient Boosting Classifier, AdaBoost Classifier (with a Decision Tree as the base classifier), KNN Classifier, Gaussian Naive Bayes and Decision Tree Classifier. The algorithms were used as implemented in Scikitlearn Python ML library [6]. For each algorithm, a randomized search on hyper parameters was performed on the training data using 2-fold cross-validation.

4. https://www.tensorflow.org/
5. https://keras.io/
6. http://scikit-learn.org/

In addition to the classical ML algorithms, we experimented with a stack of ML classifiers in order to provide a fair comparison to the novel DNN meta learning approach. The details for the optimization of the stack's parameters are thoroughly explained in our previous publication where it is experimentally shown that 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.

5.1. Evaluation results

The models were evaluated using trial-specific 10-fold cross-validation, i.e., the data segments that belong to one trial (e.g., one affective stimulus) can either belong only to the training set or only to the test set. In addition, 10% of the training data was kept as a holdout set. The holdout set is a subset of the training data which has not been seen by the base learners, is used to train the meta learners, and it is excluded from the final evaluation of the models. Depending on the target dataset, the best performing meta model on a dataset-specific holdout set was selected for the final evaluation. The results are presented in Table 2. The column "Merged" shows the accuracy of the algorithms when they are trained on the overall (merged) data. The other columns represent the accuracy of dataset-specific models.

Table 2: Accuracy for binary arousal recognition (high vs. low).

Algorithm	**Merged**	**Ascertain**	**DEAP**	**Driving**	**Cog. Load**	**Mahnob**	**Amigos**
RF	59.3	65.5	55.6	78.5	73.9	58.0	53.6
SVM	60.2	66.4	51.3	79.5	69.1	62.3	50.6
GB	59.0	64.4	53.3	75.5	76.1	60.9	54.2
AdaB	57.5	62.3	52.6	75.5	76.6	61.0	56.0
KNN	60.6	60.0	49.0	75.0	77.0	60.1	53.3
NB	60.8	59.1	53.5	66.5	80.4	62.4	45.4
DT	58.0	65.0	52.0	61.5	70.4	58.1	55.1
ML-Meta	63.0	59.0	52.5	74.4	76.3	61.8	53.8
DNN-Feat	66.2						
CNN-GSR	66.3						
DNN-Ens	70.3	64.10	52.05	78.67	76.12	83.98	66.62

From the results it can be seen that the novel DNN-ensemble method has achieved average accuracy of 70%, which is four percentage points better than the other DNN-based methods and at least seven percentage points better than the non-DNN methods. Regarding the results per dataset, on the Mahnob dataset, the DNN-ensemble method has achieved accuracy of twenty percentage points more than other methods. On the Amigos dataset, DNN-ensemble method has achieved accuracy of ten percentage points more than other methods. On th other four datasets, the DNN-ensemble method has achieved similar results as the rest of the methods.

5.2. Model analysis

The main difference between the classical ML approaches and the novel DNN ensemble was the CNN-GSR model. The DNN-Feature was trained with the same input as the classical ML approaches. For that reason, we further examined the output of the CNN-GSR network. For example, the top left graph in Figure 4 shows the normalized GSR input for the network. It can be seen that the maximum value of the signal is 0.6 which means it

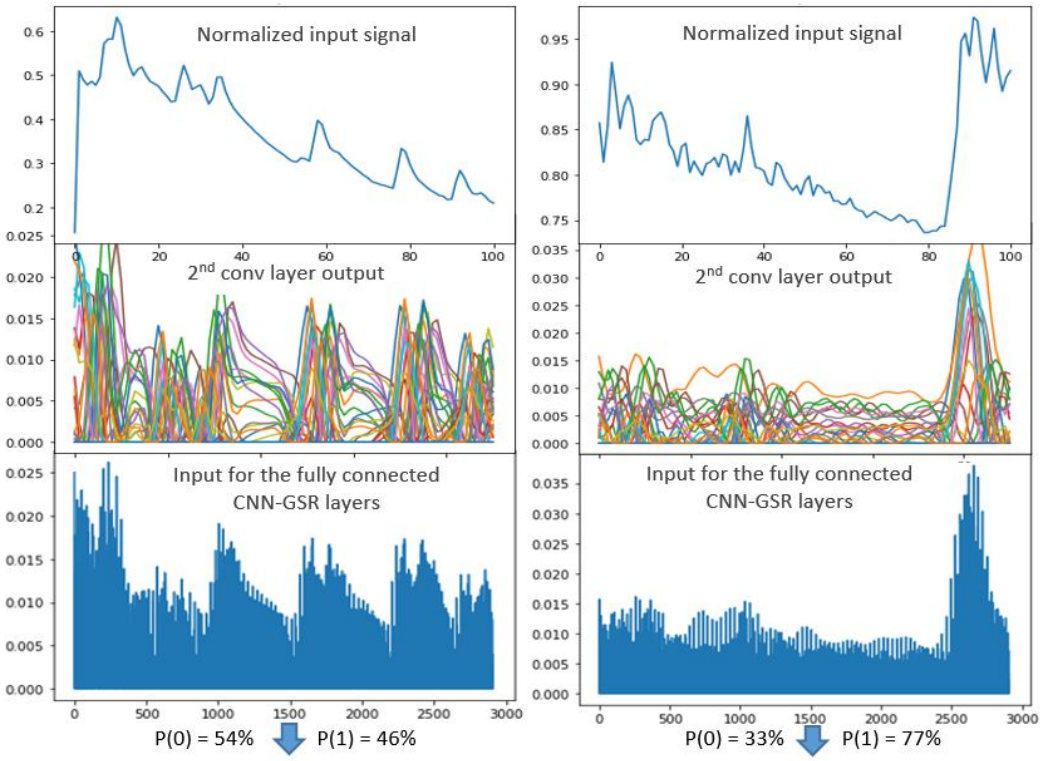

Figure 4: Example input and example output for the DNN-GSR for a low arousal (left graphs) and for high arousal (right graphs)

is near the average of the person. Also, the signal is dropping continuously, which may be a sign that the person is relaxing (low arousal). The middle left graph shows the output of the second (final) convolutional layer for the given input. This layer contains 32 different filters, thus each filter produces different representation of the signal marked with different colors on the graph. From the 32 different representations it can be seen that the CNN-GSR is emphasizing the peaks of the signal and that some of the representations are delayed, i.e., the signal's peaks are shifted in time. The bottom left grapg represents the input to the fully connected layers of the CNN-GSR. It contains 2912 input values, which correspond to the 32 different representations (CNN filters) multiplied by 91 – which is the length of the input signal after the convolutions. The bottom arrow represents the output probabilities of the CNN-GSR for the given input. In this case, the prediction is that the signal is a

"low arousal" signal with a probability of 54%. This prediction is further analyzed by the meta learner and the final output is given. Which in this case is correct. The right side of Figure 4 presents another example of the CNN-GSR network, however, in this case for a "high arousal". From the input (top-right) it can be seen that the values of the signal are over 80%, which indicates a high sweating rate. In addition to that, there is a high positive change towards the end of the signal, which may indicate affective reaction of the person. The middle-right and the top right figures show that the network emphasized the high peak of the signal. Finally, the prediction of the network (bottom-right arrow) is that the signal is a "high arousal" signal with a probability of 77%. This number is served as input to the meta learner.

6. Conclusion

The goal of this study was to improve the performance in emotion recognition based on learning from semantically similar, yet technically quite heterogeneous data. We proposed a novel DNN ensemble method and compared it against seven "flat" ML algorithms and one advanced ML stacking scheme. At least for the tested six domains, it turned out that by using DNN methods and by merging different datasets the accuracy of the affect recognition increased. The ensemble DNN method was best able to combine the abstract knowledge encompassing several domains and enrich it with the special knowledge for each domain. To a certain point, this resembles human learning: we are able to capture general, abstract common knowledge and enrich it with specialized knowledge for a specific task.

The preprocessing method used for translating different datasets into a common spectro-temporal space was a prerequisite. Once the datasets were transformed to the common spectro-termporal space, the novel DNN meta learning was able to improve upon the performance of the classical flat ML approaches by utilizing the knowledge from the different sources.

By examining the output of the CNN-GSR network, it was noted that the CNN has taught itself that peaks of the GSR signal contain information regarding arousal, thus the network developed filters to stress those parts of the signals (see Figure 4). This is in line with many physiological van Dooren et al. (2012) and affective computing Healey and Picard (2005) studies which analyze Skin Conductance Responses (SCR) s one of the main features for affect recognition.

In this paper, we focused on recognizing the arousal, i.e., the dimension that represents the intensity of the emotions. For future work we plan to extend the approach to the other two dimensions (valence and dominance). Additionally, we plan to examine the behavior of the DNNs on a completely new domain, which has not been included in the train phase.

References

J. Abdon Miranda-Correa, M. Khomami Abadi, N. Sebe, and I. Patras. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *ArXiv e-prints*, February 2017.

Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning representations from eeg with deep recurrent-convolutional neural networks.

CoRR, abs/1511.06448, 2015. URL http://dblp.uni-trier.de/db/journals/corr/corr1511.html#BashivanRYC15.

R Castaldo, P Melillo, U Bracale, M Caserta, M Triassi, and L Pecchia. Biomedical Signal Processing and Control Acute mental stress assessment via short term HRV analysis in healthy adults : A systematic review with meta-analysis. *Biomedical Signal Processing and Control*, 18:370–377, 2015. ISSN 1746-8094. doi: 10.1016/j.bspc.2015.02.012. URL http://dx.doi.org/10.1016/j.bspc.2015.02.012.

Maurizio Garbarino, Matteo Lai, Simone Tognetti, Rosalind Picard, and Daniel Bender. Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. *Proceedings of the 4th International Conference on Wireless Mobile Communication and Healthcare - "Transforming healthcare through innovations in mobile and wireless technologies"*, pages 3–6, 2014. doi: 10.4108/icst.mobihealth.2014.257418. URL http://eudl.eu/doi/10.4108/icst.mobihealth.2014.257418.

Martin Gjoreski, Mitja Lustrek, Matjaz Gams, and Hristijan Gjoreski. Monitoring stress with a wrist device using context. *Journal of Biomedical Informatics*, 73:159–170, 2017. URL http://dblp.uni-trier.de/db/journals/jbi/jbi73.html#GjoreskiLGG17.

Martin Gjoreski, Blagoj Mitrevski, Mitja Luštrek, and Matjaž Gams. An Inter-Domain Study For Arousal Recognition From Physiological Signals. 42:61–68, 2018. URL http://www.informatica.si/index.php/informatica/article/view/2232/1157,.

J. A. Healey and R. W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2): 156–166, June 2005. ISSN 1524-9050. doi: 10.1109/TITS.2005.848368.

Daniela Iacoviello, Andrea Petracca, Matteo Spezialetti, and Giuseppe Placidi. A real-time classification algorithm for eeg-based bci driven by self-induced emotions. *Computer Methods and Programs in Biomedicine*, 122(3):293–303, 2015. URL http://dblp.uni-trier.de/db/journals/cmpb/cmpb122.html#IacovielloPSP15.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. pages 448–456, 2015. URL http://dl.acm.org/citation.cfm?id=3045118.3045167.

Mahdi Khezri, Seyed Mohammad P. Firoozabadi, and Ahmad-Reza Sharafat. Reliable emotion recognition system based on dynamic adaptive fusion of forehead biopotentials and physiological signals. *Computer Methods and Programs in Biomedicine*, 122(2): 149–164, 2015. URL http://dblp.uni-trier.de/db/journals/cmpb/cmpb122.html#KhezriFS15.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

S. Koelstra, C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis ;using physiological signals. *IEEE*

Transactions on Affective Computing, 3(1):18–31, Jan 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.15.

Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal emotion recognition using multimodal deep learning. *CoRR*, abs/1602.08225, 2016. URL http://dblp.uni-trier.de/db/journals/corr/corr1602.html#LiuZL16.

N.R. Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39:447–462, 1976.

Hector P. Martinez, Yoshua Bengio, and Georgios Yannakakis. Learning deep physiological models of affect. *IEEE Computational Intelligence Magazine*, 8(2):20–33, 2013. ISSN 1556603X. doi: 10.1109/MCI.2013.2247823.

Raja Majid Mehmood and Hyo Jong Lee. A novel feature extraction method based on late positive potential for emotion recognition in human brain signal patterns. *Computers Electrical Engineering*, 53:444–457, 2016. URL http://dblp.uni-trier.de/db/journals/cee/cee53.html#MehmoodL16.

Rosalind W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997. ISBN 0-262-16170-2.

J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980. ISSN 0022-3514.

Stefan Schneegass, Bastian Pfleging, Nora Broy, Albrecht Schmidt, and Frederik Heinrich. A data set of real world driving to assess driver workload. *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '13*, pages 150–157, 2013. doi: 10.1145/2516540.2516561. URL http://dl.acm.org/citation.cfm?doid=2516540.2516561.

Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affective Computing*, 3(1):42–55, 2012. URL http://dblp.uni-trier.de/db/journals/taffco/taffco3.html#SoleymaniLPP12.

R. Subramanian, J. Wache, M. Abadi, R. Vieriu, S. Winkler, and N. Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, pages 1–1, 2017. ISSN 1949-3045. doi: 10.1109/TAFFC.2016.2625250.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn W. Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. pages 5200–5204, 2016. URL http://dblp.uni-trier.de/db/conf/icassp/icassp2016.html#TrigeorgisRBMNS16.

Marieke van Dooren, J.J.G. (Gert-Jan) de Vries, and Joris H. Janssen. Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology Behavior*, 106(2):298 – 304, 2012. ISSN 0031-9384. doi: https://doi.org/10.1016/

j.physbeh.2012.01.020. URL http://www.sciencedirect.com/science/article/pii/S0031938412000613.

Gyanendra K. Verma and Uma Shanker Tiwary. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage*, 102:162–172, 2014. URL http://dblp.uni-trier.de/db/journals/neuroimage/neuroimage102.html#VermaT14.

Zhong Yin, Mengyuan Zhao, Yongxiong Wang, Jingdong Yang, and Jianhua Zhang. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer Methods and Programs in Biomedicine*, 140:93–110, 2017. URL http://dblp.uni-trier.de/db/journals/cmpb/cmpb140.html#YinZWYZ17.