Deep Affect Recognition from R-R Intervals

Martin Gjoreski
Department of Intelligent Systems, Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
martin.gjoreski@ijs.si

Hristijan Gjoreski
University of Sussex
Brighton, United Kingdom

Mitja Luštrek
Matjaž Gams
Department of Intelligent Systems, Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Abstract
Affect recognition is an important task in ubiquitous computing, in particular in health and human-computer interaction. In the former, it contributes to the timely detection and treatment of emotional and mental disorders, and in the latter, it enables indigenous interaction and enhanced user experience. We present an inter-domain study for affect recognition on seven different datasets, recorded with six different sensors, three different sensor placements, 211 subjects and nearly 1000 hours of labelled data. The datasets are processed and translated into a common spectro-temporal space. The data represented in the common spectro-temporal space is used to train a deep neural network (DNN) for arousal recognition that benefits from the large amounts of data even when the data are heterogeneous (i.e., different sensors and different datasets). The DNN approach outperforms the classical machine-learning approaches in six out of seven datasets.

UbiComp/ISWC'17 Adjunct , September 11–15, 2017, Maui, HI, USA

ACM 978-1-4503-5190-4/17/09…$15.00
https://doi.org/10.1145/3123024.3125608

Author Keywords
Arousal recognition; Affect; Deep neural networks; Machine learning; Transfer learning; Stress, Emotions

ACM Classification Keywords
J.3 Computer Applications: Health

Introduction
It has been two decades since Rosalind Picard introduced the field of affective computing [1] and yet modeling affective states remains a challenging task. It is, however, an important one, both in the domain of human-computer interaction (HCI) and health. In the former, it enables a more natural interaction and better user experience. In the latter, it contributes to the

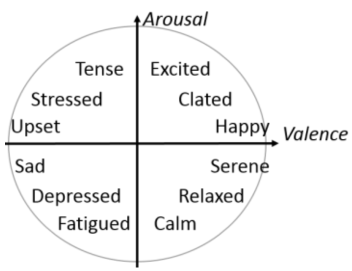

Figure 1. Circumplex model of affect. The model maps affective states in a 2D space of Arousal and Valence [3].

timely detection and treatment of emotional and mental disorders such as depression, bipolar disorders and posttraumatic stress disorder (PTSD). In 2013, the cost of work-related depression in Europe was estimated to €617 billion annually. 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) [2].

Affective states are complex and usually have fuzzy boundaries. The model that deals with the vague definitions and fuzzy boundaries of affect is the circumplex model of affect (Figure 1). The model maps the affective states into a 2D space of arousal and valence [3]. This model has been widely used in HCI studies for annotating affective states [4] [5] [6]. The use of the same annotating model allows for an inter-study analysis, which we exploit in our work. In this paper we examine arousal recognition from physiological data captured via chest-worn Electrocardiography (ECG) sensors, blood volume pulse (BVP) sensor placed on the finger or wrist-worn pulse oximeter (PPG) sensor. The data belongs to seven publicly available datasets for affect recognition. Overall, nearly 1000 hours of arousal-labelled data that belong to 211 subjects (181 different subjects, 140 males and 71 females) was analyzed. To the best of our knowledge, this is the first study on affect recognition performed on such a big amount of data.

The data comes from seven different studies and six different sensors. This introduces the problem of inter-domain learning, to which ML techniques are sensitive. To overcome this problem, we exploit two solutions. First, we use pre-processing techniques to translate the data to a common spectro-temporal space of R-R intervals and Lomb-Scargle periodogram [7], regardless of the sensor. Second, we examine transfer learning between datasets. More specifically, we examine the performance of pre-trained DNN models, i.e., models trained on other datasets and adapted to a new dataset, and new DNN models, i.e., models trained only on the new dataset. The transfer learning increases the overall amount of training data and may decrease the training time of the DNN [8].

Highlights of the study: (a) First inter-domain study for affect recognition dealing with seven different datasets recorded with six different sensors, three different sensor placements, 211 subjects (181 different subjects) and nearly 1000 hours of labelled data; (b) Pre-processing method 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; (c) DNN approach for arousal recognition that benefits from large amounts of data even when the data are heterogeneous (i.e., different sensors and different datasets), and outperforms the classical ML approach.

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 [9], video analysis [10], 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.

Table 1. Data information (number of subjects, mean age, number of trials per subject, mean duration of each trial, duration of data per subject - in seconds, and overall duration)

	Subjects (M + F)	Mean age	Trials	Duration in seconds		
				µ trial	Per subject	Overall data
ASCERTAIN	58 (37+21)	31	36	80	2880	167040
DEAP	32 (16+16)	26.9	40	60	2400	76800
DECAF_Movie	30 (16+14)	27.3	38	60	2280	68400
DECAF_Music	30 (16+14)	27.3	40	60	2400	72000
Driving	10 (7+3)	35.6	1	1800	1800	18000
Cognitive	21 (21+/)	28	2	2400	4800	100800
Mahnob	30 (13+17)	26	40	80	3200	96000
Overall	211	28.9 (µ)	197	648.6	19760	599040

In general, the methods based on EEG data outperform the methods based on other data [4] [5], 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 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 [11] and Microsoft Band [12]). 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 [13]. Khezri et al. used EEG combined with GSR to recognize six basic emotions via K-nearest neighbors (KNN) classifiers [14]. Verma et al. [15] developed an ensemble approach using EEG, electromyography (EMG), ECG, GSR, and EOG. Mehmood and Lee used independent component analysis to extract emotional indicators from EEG, EMG, GSR, ECG, and (effective refractory period) ERP [16]. Mikuckas et al. [17] 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.

Recently, the use of deep learning for affect recognition has become popular too. Liu et al. [18] presented a deep learning approach for emotion recognition using EEG data and eye blink data. They experimented on two different datasets, DEAP and SEED dataset [19]. The SEED dataset contains only EEG signals, thus it was not included in our study. Similarly, Bashivan et al. [20] presented an approach for learning representations from EEG signal with deep recurrent-convolutional neural networks. Yin et al. presented an approach for 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 [21]. In contrast to the EEG based methods for affect recognition, Martinez et al. [22] has presented a DNN method for affect recognition from GSR and BVP data.

The related work shows that – similarly to many other fields – deep learning can outperforms classical ML in affect recognition. However, the work done so far could not take full advantage of deep learning because training a DNN models requires a large amount of data, which is a problem in the field of affect recognition where datasets are usually small – not nearly the size of the datasets used in other fields (e.g. ImageNet contains 1.2 million images). The challenge may be even bigger if simpler (and more practical) hardware is used that has only one sensor modality. To overcome this challenge, we explore inter-dataset transfer learning. Transfer learning has been proven in other fields to improve the accuracy of the models or at least to improve the training speed (e.g., in computer vision [23] and activity recognition [8]).

Data

At the beginning of our study, a dataset overview was performed to find available affective datasets. We were able to target seven different datasets: ASCERTAIN [4], DEAP [5], DECAF Movies [6], DECAF Music [6], Driving workload dataset [24], Cognitive load dataset [25] [26] and MAHHNOB [27]. General information for each dataset is presented in Table 1. The table presents the number of subjects per dataset, the mean age, the

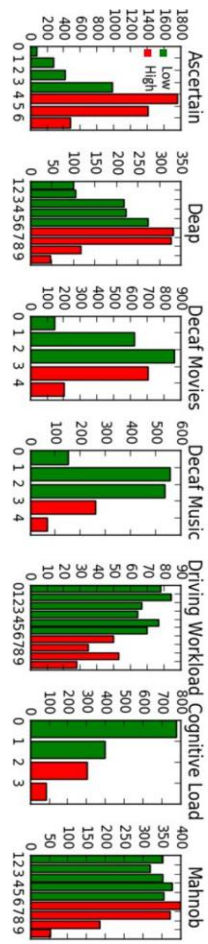

Figure 2. Label distribution per dataset. On the x-axis is the arousal level as labeled in the original dataset. The color represent the arousal level (low/high) used in our study.

number of trials per subject, the mean duration of each trial, the duration of data per subject (in seconds) and the overall duration of the data. The five datasets, ASCERTAIN, DEAP, DECAF Movies, DECAF Music, and MAHHNOB 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 7-point arousal scale, whereas the DEAP dataset used 9-point arousal scale (1 is very low, and 9 is very high). From the both scales, we split the labels in the middle, which is the same split used in the original studies [4] [5]. Similar step was performed for the datasets DECAF Movies, DECAF Music, and MAHHNOB.

The two datasets, Driving workload and Cognitive load, did not contain labels for subjective arousal level. The Driving workload dataset contained labels from subjective ratings for a workload during driving sessions. For this dataset, we presumed that increased workload corresponds to increased arousal. Thus, we used the workload ratings as an arousal ratings. The split for high arousal was put on 60%. Similarly, the cognitive load dataset contained labels for subjective stress level during stress inducing cognitive load tasks (mathematical equations). The subjective scale was from 0 to 3 (no stress, low, medium and high stress). We put the limit for high arousal on 2 (medium stress).

Figure 2 presents the label distribution for each dataset in the original study and the label distribution after the translating the labels to low/high arousal. It can be seen that for all dataset except for the ASCERTAIN dataset, the majority label is "low arousal". Besides the labels, we used the ECG data from the ASCERTAIN, DECAF Movies, DECAF Music, MAHHNOB and Driving workload database. We used the BVP data from the DEAP database (it does not contain an ECG data) and we used the R-R data from the Cognitive load dataset, which also does not contain an ECG data.

Methods

We tested two approaches for arousal recognition: DNN and classical ML. Before the tests, a pre-processing method is applied for translating the datasets to a common spectro-temporal space. The two approaches and the pre-processing method are described in the following subsections.

Pre-processing

The pre-processing method is essential and allows the merging of the seven different datasets. It translates the physiological signals (ECG or BVP) to R-R intervals and performs temporal and spectral. First, a peak detection algorithm is applied as suggested by Negri [28]. The parameter "minimum distance between peaks" was set to the half of the signal sampling rate, which guides the algorithm to detect peaks that are more than half a second apart. Figure 3 presents an example ECG signal with detected peaks. Similarly, peaks are detected from the BVP signal. Then, the preprocessing splits into two, i.e., temporal and spectral analysis.

Temporal analysis, i.e., calculating the time distance between the detected peaks represent the R-R intervals. First, each R-R signal is filtered using median filter. The median removes the R-R intervals that fall out of the interval [ɑ*median, (2- ɑ) median], where the median is the median of the R-R signal. The parameter ɑ was experimentally set to 0.7. After the median filter, person specific winsorization [29] is

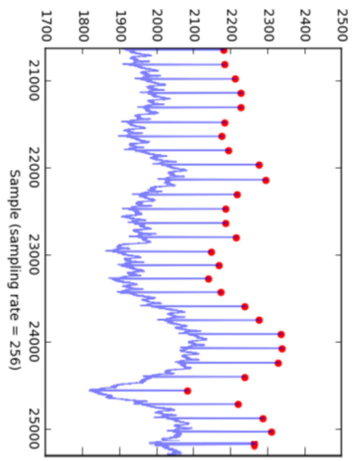

Figure 3. ECG signal and detected R-R intervals. ASCERTAIN dataset [4], Subject 1, Video 29.

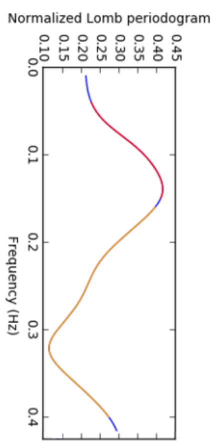

Figure 4. R-R signal represented as a time-series. ASCERTAIN dataset [4], Subject 1, Video 29

performed with the threshold parameter of 3, to remove outlier R-R intervals. The filtered R-R segments are used as input to the DNN. For the standard ML, the filtered R-R segments are used to calculate time-domain HRV features.

From the filtered R-R signals, periodogram (Figure 4) is calculated using the Lomb-Scargle algorithm developed by Lomb and further analyzed by Scargle [7][30].The Lomb-Scargle algorithm is used for spectral analysis of unequally spaced data (as are the R-R intervals). The Lomb-Scargle peridograms are used as input to the DNN. For the standard ML, the Lomb-Scargle peridograms are used to calculate frequency-domain HRV features. The red portion of the periodogram in Figure 4 is the low frequencies (lf) segment and the orange segment is the high frequencies (hf) segment.

Deep Neural Network

We used a fully connected DNN with seven hidden layers. Each layer employ rectified linear units (ReLUs). To avoid overfitting, L2 regularization and dropout was methods were used. The keep probability of the dropout was set to 0.75. The training is fully supervised, by backpropagating the gradients through all layers. The parameters are optimized by minimizing the crossentropy loss function using ADAM optimizer [31]. All models were trained with a learning rate of 10^{-4}. The batch size was set to 256 when one dataset was used, and 512 when all datasets were used for training. The output of the model is obtained from the final layer with a softmax activation function yielding a class probability distribution. The neural network was implemented using Tensorflow [32].

Classical ML Methods

For training the classical ML classifiers, a typical approach was used where the input to the ML algorithms are features extracted using HRV analysis on the filtered R-R intervals. Overall thirteen features were extracted: meanHR, meanRR, sdnn, sdsd, rmssd, pnn20, pnn50, sd1, sd2, sd1/sd2, lf, hf, lf/hf [33].

Experiments were performed with four different ML algorithms: Random Forest, Support Vector Machine, Gradient Boosting Classifier, and AdaBoost Classifier. The algorithms were used as implemented in the Scikit-learn, the Python ML library [34]. For each algorithm, randomized search on hyper parameters was performed on the training data using 2-fold validation. The hyper parameter tuning contributes towards farer comparison of the standard ML algorithms to the DNN.

Experimental Results

Two types of experiments were performed. In the first experiments we analyze the performance of the proposed DNN when transfer learning is used. The second type of experiments were performed to compare the performance of the proposed DNN to classical ML approaches. The details for each experiment are presented in the following subsections.

DNN experiments

Experiments were performed to compare the performance of a pre-trained DNN with a new DNN. The evaluation was done using the following steps: One dataset was picked as a domain dataset. The pre-trained DNN was pre-trained on the remaining six datasets for 1000 epochs (empirically chosen). After the pre-training, the training was finished on the domain dataset. On the domain dataset, leave-one-

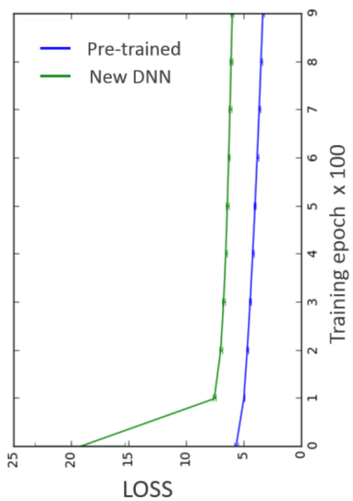

Figure 5. Training Loss for a pre-trained DNN and not pre-trained DNN (new DNN).

trial-out evaluation was performed. The average LOSS from these experiments is presented in Figure 5. The figure shows that the pre-trained DNN requires shorter time for training compared to the new DNN and it achieves lower LOSS. In these experiments, we also tried leave-one-dataset-out evaluation approach, but the results were not significantly better than a majority classifier.

DNN vs classical ML

In these experiments, we compare the performance of the proposed DNN with a classical ML algorithms. We used leave-one-trial out evaluation technique. For the classical ML algorithms we report results for person-specific models which performed better than dataset-specific models. The DNN was pre-trained on the data from all datasets except the domain dataset and it was evaluated using leave-one-trial-out on the domain-dataset. The reaming data from the subject to which belonged the testing trail was used as a validation data to tune the DNN. The results are presented in Figure 6. The results show that on average, the DNN outperforms the traditional ML algorithms. In particular, it achieves the best accuracy in 6 out of 7 datasets. Additionally, the standard deviation for the DNN is significantly lower compared to the other methods, which suggests that the DNN achieves much more stable results for the different folds.

Finally, we present visualization of the DNN models built using the leave-one-trial-out evaluation. Figure 7 presents a t-distributed stochastic neighbor embedding (t-SNE) [35] visualization of the 7-th DNN layer. This is a dimensionality reduction technique that is used for visualizing high-dimensional data. It models each high-dimensional data point by a two-dimensional point in a way that similar objects are modeled by nearby points and dissimilar objects by distant points. We first ran t-SNE with dataset-specific input, resulting in the first seven plots. In addition, we run t-SNE with all the data merged together and the output is presented in the last plot. In general, the more isolated the data is (the islands in the plots) and the more pure each isolated island is (the same class color) the better.

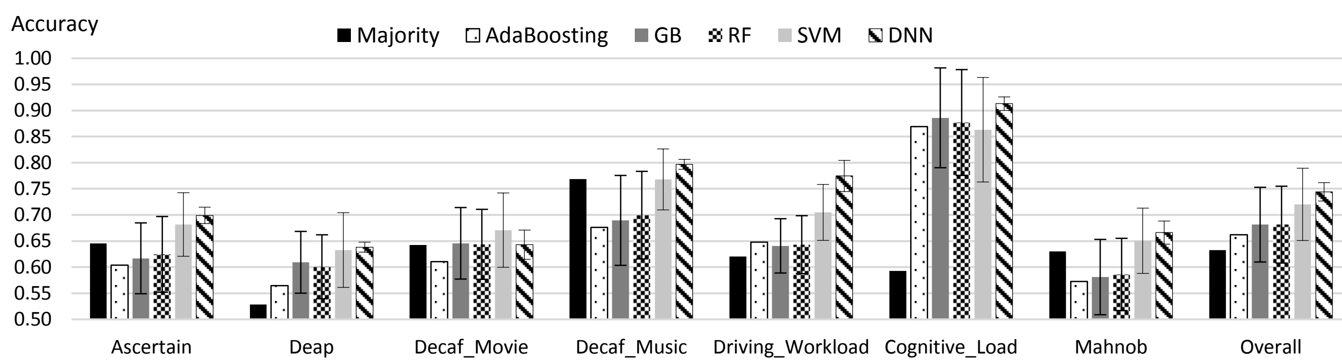

Figure 6. Performance comparison of DNN with classical ML algorithms.

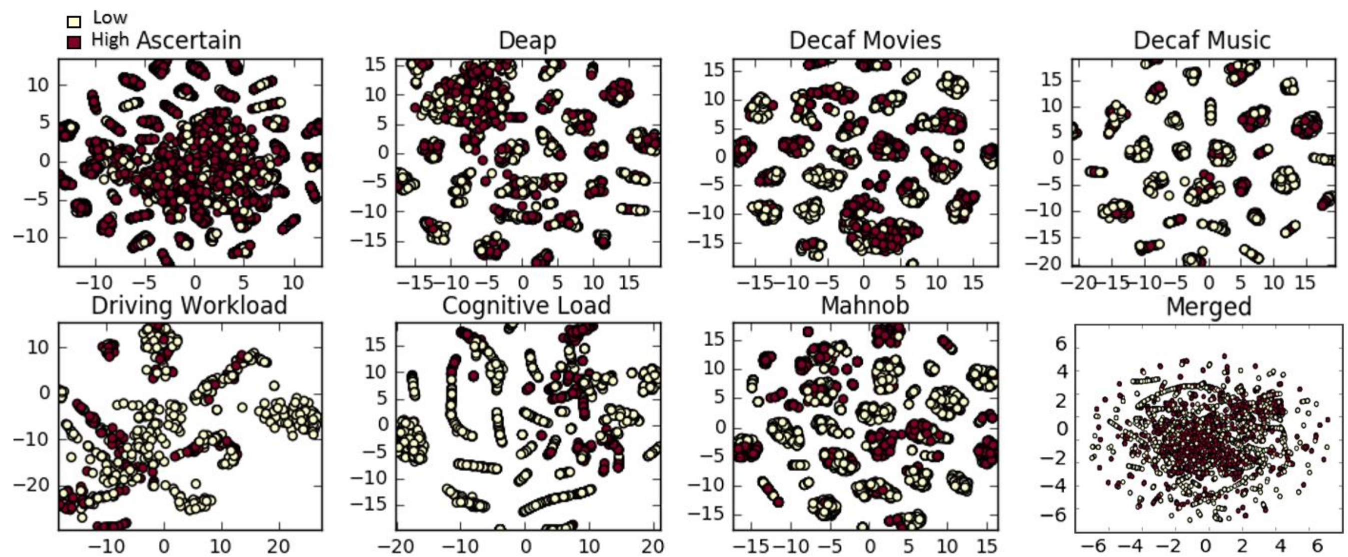

Figure 7. t-SNE visualization of the DNN weights of the final (before output) layer for each dataset (first seven plots) and merged dataset (last plot).

Discussion and Conclusion

The two central goals of our study were to improve generality and quality in affect recognition. At least for the tested seven domains, it turned out that these two properties support each other, i.e. by using DNN method and by merging different datasets the accuracy of affect recognition increased, even though the datasets are heterogeneous. The pre-processing method used for translating different datasets into a common spectro-temporal space was a prerequisite, but not enough in itself – neither with deep nor with classical ML. It turned out that even though nominally compatible arousal labels could be assigned to all the data, the experiments in which the datasets were recorded were sufficiently different that one arousal was different from another, rendering learning unsuccessful. However, when the novel DNN was pre-trained on all datasets but one, the training could successfully be finalized on the target dataset: the resulting DNN was in general more accurate and trained more quickly than a DNN trained on the target dataset only. Such an approach cannot be easily replicated with the existing classical ML algorithms, so it was only compared to classical person-specific models trained on the target dataset.

In general, the achieved accuracy is not on a satisfactory level, however the presented approach for merging seven different datasets opens a huge exploration space for future studies on affect recognition. We strongly believe that the approach may be expanded by adding the data from the GSR sensors, since five of the seven datasets contain data from GSR sensors). Thus, we plan to extend our inter-dataset study for the other sensor modalities.

Finally, the t-SNE visualization of the final DNN implies that the DNN mainly has learned person-specific models. One reason for this may be the evaluation technique. More specifically, leave-one-trial-out evaluation technique was used, and the remaining data from the subject to which belonged the testing trail was used as a validation data to tune the DNN. Additional tests should be performed (e.g., leave-one-subject-out or 10-fold evaluation) to confirm this conclusion.

References

1. R. Picard. Affective Computing. Cambridge, MA: MIT Press, 1997.
2. Depression cost: http://ec.europa.eu/health//sites/health/files/mental_health/docs/matrix_economic_analysis_mh_promotion_en.pdf, [Accessed 27.03.2017].
3. J. A. Russell. A circumplex model of affect. Journal of Personality and Social Psychology, 1980.
4. 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.
5. 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.
6. 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.
7. N.R. Lomb. Least-squares frequency analysis of unequally spaced data. Astrophysics and Space Science, vol 39, pp. 447-462, 1976
8. F. J. O. Morales, D. Roggen. Deep Convolutional Feature Transfer Across Mobile Activity Recognition Domains, Sensor Modalities and Locations. Proceedings of the 2016 ACM International Symposium on Wearable Computers (ISWC), 2016.
9. G. Trigeorgis et al.Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network,” Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.
10. 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.
11. 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.
12. Microsoft band. https://www.microsoft.com/microsoft-band/en-us
13. 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.
14. M. Khezria, M.Firoozabadib, A. R. Sharafata. Reliable emotion recognition system based on dynamic adaptive fusion of forehead biopotentials and physiological signals.
15. G. K. Verma, U. S. Tiwary. Multimodal fusion framework: A multiresolution approach for emotion

classification and recognition from physiological signals. NeuroImage, 2014.
16. 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.
17. 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.
18. 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.
19. W-L. Zheng, B-L Lu. A multimodal approach to estimating vigilance using EEG and forehead EOG. Journal of Neural Engineering, 2017.
20. P. Bashivan, I. Rish, M.Yeasin, N. Codella. Learning Representations From Eeg With Deep Recurrent-Convolutional Neural Networks. Online. Available at: https://arxiv.org/abs/1511.06448.
21. 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.
22. H.P. Martínez, Y. Bengio, G. N. Yannakakis. Learning Deep Physiological Models of Affect. IEEE Computational intelligence magazine, 2013.
23. K.Weiss, T. M. Khoshgoftaar, D. Wang. A survey of transfer learning. Journal of Big Data, 2016.
24. 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.
25. M. Gjoreski, M. Luštrek, M. Gams, H. Gjoreski. Monitoring stress with a wrist device using context. Journal of Biomedical Informatics, 2017, in press.
26. 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.
27. M. Soleymani, T.Pun. A Multimodal Database for Affect Recognition and Implicit Tagging, IEEE Transactions On Affective Computing, 2012.
28. L. H. Negri. Peak detection algorithm. Python Implementation. Online. Available at: http://pythonhosted.org/PeakUtils/.
29. M. Wu, PhD thesis. Michigan State University; 2006. Trimmed and Winsorized Eestimators.
30. 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.
31. D. P. Kingma, J. Ba. Adam: A Method for Stochastic Optimization, http://arxiv.org/abs/1412.6980, 2014.
32. Tensorflow. Online. Available at: https://www.tensorflow.org/
33. 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.
34. Scikit-learn, Python machine-learning library http://scikit-learn.org/dev/_downloads/scikit-learn-docs.pdf
35. 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.