The Cold Start Problem and Per-Group Personalization in Real-Life Emotion Recognition With Wearables.

Stanisław Saganowski[1*], Dominika Kunc[1], Bartosz Perz[1], Joanna Komoszyńska[1], Maciej Behnke[1,2], Przemysław Kazienko[1]

[1]*Department of Artificial Intelligence, Wrocław University of Science and Technology, Wrocław, Poland*
[2]*Faculty of Psychology and Cognitive Science, Adam Mickiewicz University, Poznan, Poland*

[*]stanislaw.saganowski@pwr.edu.pl

2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops) | 978-1-6654-1647-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/PerComWorkshops53856.2022.9767233

***Abstract*—Emotion recognition in real life from physiological signals provided by wrist worn devices still remains a great challenge especially due to difficulties with gathering annotated emotional events. For that purpose, we suggest building pre-trained machine learning models capable of detecting intense emotional states. This work aims to explore the cold start problem, where no data from the target subjects (users) are available at the beginning of the experiment to train the reasoning model. To address this issue, we investigate the potential of per-group personalization and the amount of data needed to perform it. Our results on real-life data indicate that even a week's worth of personalized data improves the model performance.**

***Index Terms*—emotion recognition, field studies, personalization, cold start, physiological signals, smartwatch, Emognition**

I. Introduction and Related Work

For the last century, psychologists have been using physiological responses to affective stimuli to broaden the understanding of human emotions. Drawing on psychologists' accumulative work, scientists from the affective computing domain started using psychophysiological signals to develop algorithms to detect, process, and adapt to others' emotions. To allow machines to learn about specific emotions, researchers must acquire extensive and comprehensive datasets that offer abundant emotions and diverse physiological signals collected in an ecologically valid context, i.e., real life. However, the field of emotion recognition from psychophysiological signals has been dominated by laboratory studies in which emotions are elicited with standardized affect induction procedures. This limitation has recently been overcome by researchers collecting everyday life emotions with wearables [1], [2] and Experience Sampling Methods - ESM [3] (also referred to as a daily diary method, or Ecological Momentary Assessment - EMA [4]–[6]).

Using embedded sensors from popular wearables like smartwatches or wrist bands makes it possible to measure the behavioral and physiological components of emotions [1], [7]. Only a few studies were trying to recognize real-life emotions, i.e., research in the field, especially [2], [8], [9] lasting a dozen days, on pupils in the classroom [10], on workers in the factory [11], or [12]–[14] focusing primarily on mood. Except for [2], [15], these studies did not try to recognize emotions in particular points in time but rather averaged over a longer period.

Some other researchers tried to distinguish emotion in only one specific shorter life context, e.g., while walking along a specific route in the city for a few dozen minutes [16]–[18], or babies playing in the limited area [19]. Schmidt et al. provided some hints for such studies in the wild [20].

The crucial still open question is how to find the real-life moments in which individuals experience short noteworthy emotions. The ESM provides high ecological validity of the repeated in-the-moment experience measurement, in which participants receive the measurements' notifications in a semi-random design. However, ESM can be further improved with the recent developments in affective computing, in which the measurement moments can be detected by physiologically or behaviorally driven pre-trained machine learning (ML) models [15], [21].

Overall, the ML models consist of the architecture/classifier and the data. It raises an additional issue - we need some initial data to train the pre-trained models. This issue is similar to the *cold start problem* commonly encountered and considered for recommender systems [22], [23]. The essence of the cold start problem is to prepare the system (model) to work for unknown users, for which we have not collected any prior data.

Nevertheless, if we possess data from earlier field studies, we can create an initial model. Unfortunately, there are no publicly available datasets gathered in the field, and the researchers have to rely on data acquired on their own. Alternatively, we can use data collected in the lab, which in recent years become more accessible [24]–[26]. However, the model trained on data captured in the controlled environment may perform poorly in real life [27].

Once we have the initial model, an interesting question arises - what should we do after running the study for a couple

This work was partially supported by National Science Centre, Poland, project no. 2020/37/B/ST6/03806; by the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology; by the Polish Ministry of Education and Science, National Information Processing Institute — the CLARIN-PL Project.

of weeks and when a sufficient number of samples is collected. Should we just add new data before retraining the model, replace some old cases, or create a new model trained only on the new data? The decision is relatively easy when the previous and new studies are similar in setup – the same participants, assessments, and apparatus. However, it is common to run a new study/iteration with new participants and/or slightly change (improve) the setup based on the feedback from the previous studies. In such a case, the model trained on the data from the previous study might not perform well because of a different set of participants (emotion recognition models are known to have poor generalization ability [28], [29]), different setup, or in general due to concept drift [30], [31].

In this work we investigate four scenarios of retraining and replacing model once the sufficient number of samples are collected: Scenario S1 – utilizing *old* model; S2 – replacing some *old* data with the *new* one; S3 – training model on *new* samples only; S4 – retraining model using all available data (*old* and *new*). All our experiments in real-life show that adding new knowledge improves the model's performance, but the best results were achieved by the model made using only new data.

To some extent, this is similar to model personalization. Except we personalize the model per group of participants rather than per a single participant. There were several attempts of per-participant model personalization in the field studies; however, they were unsuccessful due to the low number of per-person samples [28]. Per-group personalization can mitigate this problem.

II. Study Setup and Data

A. The Emognition Framework

The Emognition system [15] includes a mobile Android application with an embedded pre-trained ML model, a smartwatch application recording physiological signals, and a back-end server storing all data. The smartwatch used in the framework is Samsung Galaxy Watch 3, and the smartphones are Android-based devices owned by the participants. The connection between smartwatch and smartphone is handled by the Bluetooth Low Energy module. The 45mm version of the smartwatch, equipped with a 330mAh battery, can record up to 14 hours of physiological data before running out of power, while the smaller one, 41mm version with a 240mAh battery, can work for up to nine hours. Physiological data are recorded continuously and noninvasively. The smartwatch provides raw blood volume pulse (BVP) sampled at 25 Hz, heart rate (HR) sampled at 12.5 Hz, RR-interval (RRI) sampled at 12.5 Hz, and 3-axis accelerometer data (ACC) sampled at 50 Hz. The device provides other data: 3-axis gyroscope, 4-axis rotation, pressure, and ambient light. One hour of recording produces about 8.6 MB of compressed data. The data is transferred to the smartphone in real-time, and from there is uploaded to the back-end server every hour. The upload can also be triggered by the user.

For more details regarding the Emognition system, please refer to [15].

B. Data

In recent months, we have performed two daily life studies. The studies were alike but had a different set of participants and slightly modified self-assessment. We will refer to them as Study A and Study B.

The primary goal of Study A was to collect physiological signals during emotionally intense moments in participants' everyday lives. The collected emotionally annotated signals were then used for creating an ML model recognizing intense emotions in real-time [21]. The model was further used for more efficient data gathering in Study A and Study B. Study A involved 11 participants (four females) and lasted about seven months.

The main idea behind Study B, which is currently still in progress, is the validation of several various predictive models and further data collection. Study B involves 13 participants (six females) and is designed to last two months. In the analysis, we consider only the first four weeks of Study B and only five participants (two females) with the highest number of reported self-assessments. The changes introduced in the Emognition system in Study B include shorter self-assessment and three types of assessment triggers.

Participants' emotions were collected with brief questionnaires using ESM at quasi-random times, machine learning triggered, and self-initiated reports. First, participants were asked whether they felt intense emotions (yes/no/not sure). Based on this question, we categorized emotions as intense emotions (yes) or neutral states (no). Next, participants reported valence on a slider scale from 1 (extremely negative) to 100 (extremely positive), and arousal on a slider scale from 1 (extremely sluggish) to 100 (extremely aroused). Finally, participants had the opportunity to provide some comments as a free text.

In total, 1075 (440 intense emotions and 635 neutral states) self-reports were collected throughout both studies (Tab. I). The total participants' pool of data used in analyzes consisted of 16 participants (6 female) between the ages of 18 and 54 years (M=26.86, SD = 8.29). All participants (volunteers) provided written informed consent and received no compensation for their participation. The research was approved by and performed in accordance with guidelines and regulations of the Bioethical Committee at Wroclaw Medical University, Poland; approval no. 149/2020.

III. Experimental Scenarios

We have designed four possible scenarios to choose from once the study obtains the required number of samples to create a decision model. The scenarios are visualized in Fig. 1. One of the scenarios utilizes all available data (from both studies) to train the model, whereas the other three analyze whether it is profitable to replace the previous samples in the training set with new samples. To ensure we analyze the quality of the samples, not the quantity, Scenarios S1 to S3 consider an equal number of samples.

Scenario S1 assumes training a model on data from the previous studies only (Study A). This is a classic example

TABLE I: Distribution of the studies data.

Study/Scenario	Intense emotion	Neutral	Sum
Study A	233	449	682
Study B week 1	71	61	132
Study B week 2	55	50	105
Study B weeks 3+4	65	73	138
Scenario S1 / S2 / S3	126	111	237
Scenario S4	359	342	701
Avg per person per week			
Study A	1.8±3.0	3.4±3.2	5.1±5.2
Study B	9.6±6.6	9.2±7.2	18.8±9.5

of validating the model's generalization ability since data in the test set come from different participants than data in the training set, which is the only possible scenario at the very beginning of a new study. S1 is based on 237 samples that were drawn from the entire Study A in a way that the number of samples in each class is equal to the number of samples in weeks 1 and 2 of Study B, i.e., 126 samples of intense emotions and 111 samples of neutral state were randomly selected. The sampling was repeated five times and the results presented in the latter part of the article are the average of the five runs.

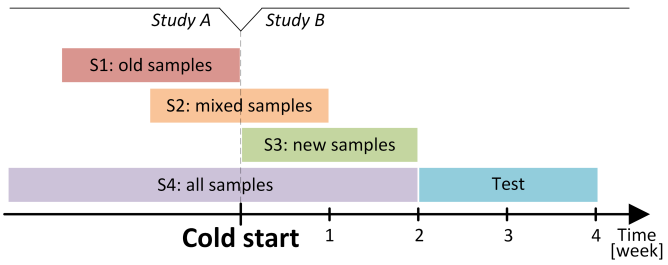

Fig. 1: Four scenarios S1–S4 of using samples to train the classification model for the field study being currently conducted.

Scenario S2 utilizes part of data from Study A and adds data from the current Study B to create a model. This scenario is possible once we obtain some new data, but the amount is still too low to create an entirely new model. S2 includes 105 samples randomly selected from Study A and 132 samples from the first week of Study B. Like in the case of Scenario S1, the sampling was repeated five times and the results were averaged.

Scenario S3, per-group personalization, considers the model trained on the new samples only. It is trained with 237 samples collected during the first and second weeks of Study B. Its advantage is the same set of participants in the training and test sets.

Scenario S4, on the other hand, makes use of all available samples, i.e., data collected in Study A (undersampled to achieve balanced data) and all the data from the first two weeks of Study B. In total, 701 samples are used to build the predictive model.

IV. Methods and Models

A. Physiological Signals

For the experiments, six types of signals were used: (1) raw BVP signal provided by the smartwatch; (2) BVP signal processed with median and band-pass Butterworth filters; (3) heart rate and (4) RR-interval, both offered by the smartwatch; (5) heart rate computed from RR-interval; (6) accelerometer data (ACC). Depending on the setup, different signals were utilized: signals 1-4 for end-to-end models (e2e); signals 1-5 (without ACC), or 1-6 (with ACC) for feature-based models.

From collected signals, we extracted windows of length 140 seconds, with the emotional event in the middle. We discarded windows with more than 10% of samples missing (compared to the expected amount, based on sampling frequency). Then, all signals were resampled using *resample* function from SciPy [32]. Next, we extracted a window of 60s around the emotional event (30s per side, event in the middle) for each signal. The window was further divided into three parts, each of length 20s. This partition was done to allow ML models to analyze the physiology before, during, and after the event, and potentially learn shorter dependencies and relations present when we experience intense emotions.

B. Features

For some experiments, it was necessary to extract features from signals. Computed features (see Tab. II) include standard statistical features like e.g., min, max, mean values of the signal, or standard deviation. Moreover, we computed differences between consecutive parts of a window for max, min, mean, std, and variance (e.g., difference between minimum values in the first and second part of a window). Furthermore, we computed features in the frequency domain, for example, minimum, maximum, or average values in the power spectrum. Additionally, for the BVP signal, we computed the mean value in low- and high-frequency power spectra. When creating a vector of features, features for all three parts of a window were concatenated, and two more date-related features were added. In total, 746 (with ACC) or 418 (without ACC) features were supplied to classifiers/architectures.

C. Models

Machine learning models used in experiments can be divided into two categories: feature-based and end-to-end. Feature-based models included AdaBoost, k-Nearest Neighbours (KNN), Random Forest, Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Fully Convolutional Network (FCN) with a kernel size of one. For experiments with end-to-end networks, we utilized FCN, FCN with additional Long Short-Term Memory (LSTM) layers (FCN-LSTM), and Residual Neural Network (ResNet). For FCN-LSTM consecutive parts of a window were first processed by FCN channels and later supplied to LSTM layers. For FCN and ResNet, all parts of a window were treated as separate channels

TABLE II: Features extracted from the physiological signals.

Signal	Domain	Features
All signals & derivatives	Statistical	min, max, min-max difference, standard deviation, variance, mean, 1st quartile, 2nd quartile, 3rd quartile, interquartile range, 1st value, last value, 1st and last values difference, skewness, kurtosis, 2nd difference mean, 2nd difference standard deviation, slope, mean difference, min difference, max difference, standard deviation difference, variance difference
	Frequency	dominant frequency, energy, max power, min power, mean power, standard deviation power
BVP	Frequency	mean of power spectrum in low frequency (0.05-0.15 Hz), mean of power spectrum in high frequency (0.16-0.4 Hz)
Time related		Day of the week (0 (Monday) - 6 (Sunday)), Hour (0-23)

(3 window parts $\times$ 4 signals = 12 channels in total). The deep learning architectures were programmed in PyTorch [33] according to an article by Dzieżyc et al. [34]. For classical machine learning algorithms, we used implementations from scikit-learn [35].

D. Model Training and Optimization

To prepare datasets for Scenarios S1, S2, and S4, data were balanced using a random sampling technique. We treated these samples as a basis for splits of data used to tune hyperparameters (5 drawings resulted in 5 splits). Each of such splits was further randomly divided into training and validation parts. For S3, which did not require balancing, data was split into five parts as well to account for differences in training and validation splits.

The best hyperparameters were chosen based on hyperparameter optimization, which was done separately for each scenario and model. Models from scikit-learn were optimized using grid search. For deep learning models, we utilized random search, as it is more efficient [36], thus more suited for the long training process. For each classifier, its hyperparameters space was tested using five-fold validation. In all cases, the best hyperparameters were chosen based on the mean F1 macro score. The best models were retrained on the whole splits and tested on the data from Study B weeks 3+4, see Fig. 1.

V. Results

The results of each scenario and model are presented in Tab. III. The highest scores for each classifier/architecture and performance measure are bolded. S3 does not have mean values as there were no random subsets of the training sets in this scenario. We consider three metrics: (1) F1 on class 1, as we aim to recognize intense emotions properly and catching all possible emotional events is more important than capturing neutral states; (2) F1 macro, to monitor the overall performance of the model in emotional and neutral states; and (3) accuracy as another overall measure.

In general, regardless of the model and feature set used, Scenario S3 performed better than other scenarios. This result shows the importance of model personalization in the emotion recognition task. The effectiveness of the predictive model gradually increases when we replace training samples from Study A (previous study) with the samples from the current Study B, see Fig. 2. This tendency is noticeable for every kind of presented approach. The best performing models were AdaBoost and SVM for feature-based classical approach, MLP with ACC for feature-based deep learning, and FCN-LSTM with ACC for e2e deep learning. The mean differences between Scenario S1 and S3, in favor of S3, are 0.09 in F1 on class 1, 0.05 in F1 macro, and 0.05 in accuracy. Particularly significant and desired is gain in F1 on class 1. A possible conclusion is that physiological traces of intense emotions are more personalized/user-dependent than physiological changes during neutral states. In several cases, models based on S4 performed better than models based on S3. This may indicate that some classifiers/architectures benefit from additional training samples, even though samples are not representative (out of the application domain). Nevertheless, in the majority of cases where the S4 model achieved higher results, the model from S3 performed within the range of the standard deviation of the S4 model.

The Friedman statistical test [37] confirmed that the model created in Scenario S3 is the top-ranked, S4 – the second-best, S2 – the third, and S1 is ranked lowest ($p = 3\mathrm{E}{-6}$). The Shaffer post-hoc multiple comparisons [38] indicated that the differences between the results of S1 and all other models are statistically significant. The difference between the results of other models, i.e., S2 vs. S3, S2 vs. S4, and S3 vs. S4, are insignificant. There is no clear indication, whether including accelerometer data improves the model. It definitely increases the complexity and computational requirements.

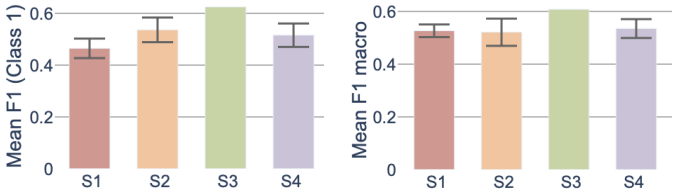

Fig. 2: Mean F1 (Class 1) and mean F1 (Macro) scores for AdaBoost classifier for all scenarios.

VI. Conclusions

Since emotional events happen in our everyday life sporadically, we should make every effort to increase the likelihood of capturing such cases with wrist-worn smartwatches. This includes personalized ML models recognizing the proper time to trigger self-assessments. However, creating personalized models requires a large number of per-person training samples, i.e., to overcome the cold start problem. Until the necessary quantity of cases is reached, we propose using an alternative, temporal solution, namely per-group personalization.

TABLE III: Results for each scenario for tested classifiers

Model	Metric	S1	S2	S3	S4
AdaBoost	F1 class 1	0.47±0.04	0.54±0.05	**0.62**	0.52±0.05
	F1 macro	0.53±0.02	0.52±0.05	**0.61**	0.53±0.04
	Accuracy	0.54±0.03	0.52±0.05	**0.61**	0.54±0.04
AdaBoost with ACC	F1 class 1	0.47±0.05	0.54±0.04	0.50	**0.56±0.03**
	F1 macro	0.53±0.05	0.51±0.04	0.48	**0.55±0.03**
	Accuracy	0.53±0.05	0.51±0.04	0.48	**0.55±0.03**
KNN	F1 class 1	0.45±0.04	0.52±0.04	**0.53**	0.49±0.03
	F1 macro	0.53±0.03	0.53±0.03	0.53	**0.55±0.01**
	Accuracy	0.54±0.03	0.53±0.03	0.53	**0.56±0.01**
KNN with ACC	F1 class 1	0.48±0.07	0.48±0.04	0.50	**0.53±0.03**
	F1 macro	0.56±0.04	0.53±0.03	0.55	**0.58±0.02**
	Accuracy	0.58±0.03	0.54±0.03	0.56	**0.58±0.02**
RandomForest	F1 class 1	0.48±0.05	**0.60±0.03**	0.58	**0.60±0.03**
	F1 macro	0.51±0.03	0.56±0.04	0.54	**0.60±0.03**
	Accuracy	0.52±0.03	0.56±0.03	0.54	**0.60±0.03**
RandomForest with ACC	F1 class 1	0.50±0.04	0.56±0.02	**0.60**	0.59±0.02
	F1 macro	0.52±0.04	0.55±0.03	0.53	**0.59±0.02**
	Accuracy	0.52±0.04	0.55±0.03	0.54	**0.59±0.02**
SVM	F1 class 1	0.49±0.04	0.56±0.05	**0.62**	0.54±0.03
	F1 macro	0.53±0.03	0.52±0.03	**0.57**	0.56±0.04
	Accuracy	0.53±0.02	0.52±0.03	**0.57**	0.56±0.04
SVM with ACC	F1 class 1	0.49±0.05	0.56±0.04	**0.61**	0.58±0.03
	F1 macro	0.53±0.03	0.54±0.04	0.57	**0.59±0.02**
	Accuracy	0.53±0.03	0.54±0.04	0.57	**0.59±0.01**
MLP	F1 class 1	0.51±0.06	0.54±0.05	**0.57**	0.48±0.03
	F1 macro	0.53±0.05	0.55±0.06	**0.57**	0.50±0.03
	Accuracy	0.53±0.05	0.55±0.06	**0.57**	0.50±0.03
MLP with ACC	F1 class 1	0.53±0.03	0.54±0.02	**0.61**	0.55±0.03
	F1 macro	0.54±0.02	0.53±0.02	**0.61**	0.56±0.03
	Accuracy	0.54±0.02	0.53±0.02	**0.61**	0.56±0.03
Resnet e2e	F1 class 1	0.44±0.11	0.50±0.15	**0.58**	0.55±0.02
	F1 macro	0.51±0.03	0.54±0.05	**0.62**	0.55±0.03
	Accuracy	0.53±0.02	0.57±0.02	**0.63**	0.55±0.03
Resnet e2e with ACC	F1 class 1	0.57±0.05	**0.62±0.04**	0.57	0.59±0.02
	F1 macro	0.52±0.06	**0.61±0.02**	0.59	0.56±0.03
	Accuracy	0.52±0.06	**0.61±0.02**	0.59	0.56±0.03
FCN e2e	F1 class 1	0.51±0.05	0.53±0.13	**0.63**	0.56±0.02
	F1 macro	0.54±0.04	0.52±0.06	**0.64**	0.58±0.01
	Accuracy	0.54±0.04	0.54±0.04	**0.64**	0.58±0.01
FCN e2e with ACC	F1 class 1	0.55±0.07	0.62±0.03	0.58	**0.62±0.01**
	F1 macro	0.52±0.05	**0.61±0.02**	0.56	0.60±0.02
	Accuracy	0.52±0.05	**0.61±0.02**	0.57	0.60±0.02
FCN-LSTM e2e	F1 class 1	0.45±0.05	0.55±0.05	0.56	**0.61±0.02**
	F1 macro	0.51±0.02	0.55±0.01	**0.60**	0.59±0.02
	Accuracy	0.52±0.02	0.56±0.01	**0.61**	0.59±0.02
FCN-LSTM e2e with ACC	F1 class 1	0.47±0.03	0.57±0.07	0.60	**0.65±0.03**
	F1 macro	0.48±0.04	0.56±0.02	**0.61**	0.61±0.02
	Accuracy	0.48±0.04	0.57±0.03	**0.62**	0.62±0.02

The analysis performed on real-life data demonstrates that adjusting the model to the group of participants (Scenario S3) improves the classification quality over the general model (Scenario S1) or partially adjusted model (Scenario S2). A large number of general samples enriched with the personal samples (Scenario S4) can improve the classification over the general or partially adjusted model (Scenario S1 and Scenario S2), however because of the large portion of the general samples, is not able to outperform the adjusted model (Scenario S3). This leads us to the conclusion that not only the quantity of the training set but mostly its quality improves the models' predictive ability. Models perform better when they are trained on data from the application domain. We can also infer that human physiology can not be easily generalized to unknown participants. Hence, the *cold start problem* is a major concern at the beginning of a new study. The solution is to collect new subjects' data and perform models personalization as soon as possible to provide better-suited predictions. An obvious approach would be to adjust models for each participant separately. We have attempted such a scenario but did not obtain satisfactory results. The most probable reason for unsuccessful per-person model personalization is the low number of per-subject samples. The number of self-assessments (annotated samples) collected per person during the first two weeks of Study B varied from 13 to 33 (avg 23.7).

Study B described in this work is still ongoing. We plan to validate the model from Scenario S3 in real life by propagating the model to the participants. Furthermore, our next step will be to enrich the prediction of the intense emotion with the valence (positive vs. negative emotion).

REFERENCES

[1] S. Saganowski, A. Dutkowiak, A. Dziadek, M. Dzieżyc, J. Komoszyńska, W. Michalska, A. Polak, M. Ujma, and P. Kazienko, "Emotion recognition using wearables: A systematic literature review-work-in-progress," in *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (EmotionAware 2020)*. IEEE, 2020, pp. 1–6.

[2] P. Schmidt, R. Dürichen, A. Reiss, K. Van Laerhoven, and T. Plötz, "Multi-target affect detection in the wild: An exploratory study," in *Proceedings of the 23rd International Symposium on Wearable Computers*, ser. ISWC '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 211–219. [Online]. Available: https://doi.org/10.1145/3341163.3347741

[3] I. Myin-Germeys and P. Kuppens, "The open handbook of experience sampling methodology," in *The Open Handbook of Experience Sampling Methodology*. The center for Research on Experience sampling and Ambulatory methods Leuven (REAL), 2021, pp. 1–311.

[4] J. M. Smyth and A. A. Stone, "Ecological momentary assessment research in behavioral medicine," *Journal of Happiness studies*, vol. 4, no. 1, pp. 35–52, 2003.

[5] M. Csikszentmihalyi and R. Larson, "Validity and reliability of the experience-sampling method," in *Flow and the foundations of positive psychology*. Springer, 2014, pp. 35–54.

[6] P. Schmidt, A. Reiss, R. Dürichen, and K. V. Laerhoven, "Wearable-based affect recognition—a review," *Sensors*, vol. 19, no. 19, p. 4079, 2019.

[7] S. Saganowski, P. Kazienko, M. Dziezyc, P. Jakimow, J. Komoszynska, W. Michalska, A. Dutkowiak, A. Polak, A. Dziadek, and M. Ujma, "Consumer wearables and affective computing for wellbeing support," in *MobiQuitous 2020 - 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ACM, 2020, p. 482–487.

[8] A. Cho, H. Lee, Y. Jo, and M. Whang, "Embodied emotion recognition based on life-logging," *Sensors*, vol. 19, no. 23, 2019.

[9] M. R. Kamdar and M. J. Wu, "Prism: A data-driven platform for monitoring mental health," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 21, pp. 333–44, 2016.

[10] M. S. Dao, D. Nguyen, D. Tien, A. Kasem *et al.*, "Healthyclassroom-a proof-of-concept study for discovering students' daily moods and classroom emotions to enhance a learning-teaching process using heterogeneous sensors," in *Proc. of the Int. Conf. on Pattern Recognition App. and Methods (ICPRAM)*. Scitepress–Science and Technology Publications, 2018.

[11] Y. Kadoya, M. Khan, S. Watanapongvanich, and P. Binnagan, "Emotional status and productivity: Evidence from the special economic zone in laos," *Sustainability*, vol. 12, p. 1544, 02 2020.

[12] A. Exler, A. Schankin, C. Klebsattel, and M. Beigl, "A wearable system for mood assessment considering smartphone features and data from mobile ecgs," in *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*, 2016, pp. 1153–1161.

[13] P. A. Gloor, A. F. Colladon, F. Grippa, P. Budner, and J. Eirich, "Aristotle said "happiness is a state of activity"—predicting mood through body sensing with smartwatches," *Journal of Systems Science and Systems Engineering*, vol. 27, no. 5, pp. 586–612, 2018.

[14] G. Valenza and E. P. Scilingo, *Autonomic Nervous System Dynamics for Mood and Emotional-State Recognition: Significant Advances in Data Acquisition, Signal Processing and Classification*. Springer, 2014.

[15] S. Saganowski, M. Behnke, J. Komoszyńska, D. Kunc, B. Perz, and P. Kazienko, "A system for collecting emotionally annotated physiological signals in daily life using wearables," in *9th International Conference on Affective Computing and Intelligent Interaction (ACII 2021)*. IEEE, 2021.

[16] S. Kim, K. A. E. Patra, A. Kim, K.-P. Lee, A. Segev, and U. Lee, "Sensors know which photos are memorable," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2017, pp. 2706–2713.

[17] E. Kanjo, E. M. Younis, and N. Sherkat, "Towards unravelling the relationship between on-body, environmental and emotion data using sensor information fusion approach," *Information Fusion*, vol. 40, pp. 18–31, 2018.

[18] E. Kanjo, E. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Information Fusion*, p. 33, 10 2019.

[19] H. Feng, H. M. Golshan, and M. H. Mahoor, "A wavelet-based approach to emotion classification using eda signals," *Expert Systems with Applications*, vol. 112, pp. 77–86, 2018.

[20] P. Schmidt, A. Reiss, R. Dürichen, and K. Van Laerhoven, "Labelling affective states" in the wild" practical guidelines and lessons learned," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 2018, pp. 654–659.

[21] M. Dzieżyc, J. Komoszyńska, S. Saganowski, M. Boruch, J. Dziwiński, K. Jabłońska, D. Kunc, and P. Kazienko, "How to catch them all? enhanced data collection for emotion recognition in the field," in *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 2021, pp. 348–351.

[22] J. Gope and S. K. Jain, "A survey on solving cold start problem in recommender systems," in *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 2017, pp. 133–138.

[23] J. Misztal-Radecka, B. Indurkhya, and A. Smywiński-Pohl, "Meta-user2vec model for addressing the user and item cold-start problem in recommender systems," *User Modeling and User-Adapted Interaction*, vol. 31, no. 2, pp. 261–286, 2021.

[24] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Scientific Data*, vol. 7, no. 1, pp. 1–16, 2020.

[25] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Scientific data*, vol. 6, no. 1, pp. 1–13, 2019.

[26] K. Roelofs, M. A. Hagenaars, and J. Stins, "Facing freeze: social threat induces bodily freeze in humans," *Psychological science*, vol. 21, no. 11, pp. 1575–1581, 2010.

[27] Y. Xu, I. Hübener, A.-K. Seipp, S. Ohly, and K. David, "From the lab to the real-world: An investigation on the influence of human movement on emotion recognition using physiological signals," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2017, pp. 345–350.

[28] G. R. Tizzano, M. Spezialetti, and S. Rossi, "A deep learning approach for mood recognition from wearable data," in *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2020, pp. 1–5.

[29] J. C. Quiroz, E. Geangu, and M. H. Yong, "Emotion recognition using smart watch sensor data: Mixed-design study," *JMIR mental health*, vol. 5, no. 3, p. e10153, 2018.

[30] L. Miranda, J. Viterbo, and F. Bernardini, "A survey on the use of machine learning methods in context-aware middlewares for human activity recognition," *Artificial Intelligence Review*, pp. 1–32, 2021.

[31] Ł. Korycki and B. Krawczyk, "Concept drift detection from multi-class imbalanced data streams," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2021, pp. 1068–1079.

[32] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[34] M. Dzieżyc, M. Gjoreski, P. Kazienko, S. Saganowski, and M. Gams, "Can we ditch feature engineering? end-to-end deep learning for affect recognition from physiological sensor data," *Sensors*, vol. 20, no. 22, 2020.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[36] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. 10, pp. 281–305, 2012. [Online]. Available: http://jmlr.org/papers/v13/bergstra12a.html

[37] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the american statistical association*, vol. 32, no. 200, pp. 675–701, 1937.

[38] J. P. Shaffer, "Modified sequentially rejective multiple test procedures," *Journal of the American Statistical Association*, vol. 81, no. 395, pp. 826–831, 1986.