2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 979-8-3503-2447-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/EMBC40787.2023.10340513

Uncovering Emotions: A Pilot Study on Classifying Moods in the Valence-Arousal Space using In-the-Wild Passive Data

Cristina G. Vazquez[1,*], Corinne Eicher[2,3,4], Reto Huber[4], Golo Kronenberg[2], Hans-Peter Landolt[3] , Erich Seifritz[2], and Giulia Da Poian[1]

***Abstract*— Mood classification from passive data promises to provide an unobtrusive way to track a person's emotions over time. In this exploratory study, we collected phone sensor data and physiological signals from 8 individuals, including 5 healthy participants and 3 depressed patients, for a maximum of 35 days. Participants were asked to answer a digital questionnaire three times daily, resulting in a total of 334 self-reported mood state samples. Gradient-boosting classification was applied to the collected passive data to categorize 4 mood states in the Valence-Energetic Arousal space. The cross-validation results showed better classification performance compared to a baseline model, which always predicts the majority class. The classifier using passive data had an area under the precision-recall curve of 0.39 (SD = 0.1) while the baseline had 0.26 (SD = 0.03), suggesting the presence of information in the collected features that support the classification process. The model identified the entropy of the heart rate and the average physical activity in the preceding 8 hours, along with the max normal-to-normal (NN) sinus beat interval and the NN low frequency-high frequency ratio during the questionnaire completion, as the most important features in its analysis. Additionally, the time range of data collection was considered a contextual factor.**

I. INTRODUCTION

Mood disorders, such as depression and anxiety, are among the most prevalent and debilitating mental health conditions worldwide [1].

Measurements of mood and emotions are widely employed in outpatient psychological and psychiatric assessments of mood disorders such as depression. However, it is currently performed with self-reported symptoms during structural clinical interviews that are infrequent and subject to observer bias [2] and variability in accuracy for recalling past symptoms [3]. This leads to diagnostic challenges and difficulties in evaluating the effectiveness of interventions. The last decade has seen an explosion in the capability of monitoring individuals via sensors in wearable devices and smartphones. These devices have shown the ability to identify objective digital biomarkers, and this capability is expected to increase in the future.

Several studies have correlated mood and depressive symptoms with social interaction measured from a combination of mobile phone sensors and apps usage [4], [5], [6]: greater depression scores have been associated with decreases in total communication and social contact (number of phone calls, text messages); and several studies have related Wi-Fi- and GPS-derived location features to depression [7], [8], [9]. The relationship between physical activity and depression has also been well-studied. Greater levels of accelerometer-based physical activity and energy expenditure were strongly associated with decreased depression rates [10]. A link between autonomous nervous system dynamics and mood swings has been suggested based on physiological parameters derived from cardiovascular activity [11]. Therefore, passive data holds great potential to provide a frequent and unobtrusive assessment of intervention outcomes.

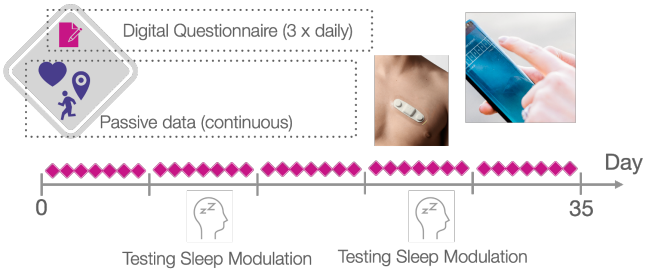

Fig. 1. Remote Assessment and Sleep Modulation (RASM) pilot study protocol and data modalities. Each of the pink diamonds represents a day in the study. The ECG patch continuously records passive data, an APP collects passive data, and digital questionnaires are filled out 3 times daily. A sleep device is worn during the nights on weeks 2 and 4.

A recent study collected passive data and self-reported mood (valence and arousal) scores, discovering statistically significant distinctions in GPS mobility, phone usage, sleep, physical activity and mood between depressed and non-depressed groups [12]. The study successfully classified the depressed and non-depressed groups using mood questionnaires and the collected passive data. A study conducted by Sükei et. al. showed that passively collected data from mobile devices could also be used for predicting emotional states in depressed patients [13].

Our research aimed at examining the use of passive data and machine learning algorithms to classify the four mood states in the Valence-Energetic Arousal space for both healthy and depressed individuals. Our objective was to determine the viability of these models for both groups and to examine the potential of wearable technology in mental health assessment. The data we gathered is representative of a clinical trial context, where tracking mood changes in

This work was supported by the Swiss National Science Foundation under Grant PZ00P2_193291 and Grant 323530_207034. This project was conducted as part of the SleepLoop Flagship of Hochschulmedizin Zurich.

*cristina.gallegovazquez@hest.ethz.ch

[1]Sensory-Motor Systems (SMS) Lab, Department of Health Sciences and Technology, ETH Zurich, Switzerland

[2]Department of Psychiatry, Psychotherapy and Psychosomatics, Psychiatric University Hospital Zurich, University of Zurich

[3] Institute of Pharmacology and Toxicology, University of Zurich

[4] Child Development Centre, University Children's Hospital, Zurich.

individuals moving from depression to recovery is essential. During the data collection process, we also tested a wearable device for sleep modulation interventions, which could introduce additional variability in the mood states and affect the features used to train the model, providing a real-world scenario.

II. MATERIAL AND METHODS

A. Data Acquisition and Dataset Characteristics

The data used in this research were collected for the Remote Assessment and Sleep Modulation (RASM) pilot study. See Fig. 1 for an overview of the study protocol. The study procedure was approved by the ETH Zurich Ethics Commission (ETHZ-EK 2021-N-153), and all participants signed an informed consent form. We asked participants to wear an electrocardiogram (ECG) patch (VivaLNK, VV330_1) on the chest for the entire study duration of 35 days. The patch recorded ECG at 128 Hz and had an embedded three-axis accelerometer capturing XYZ raw acceleration at 25 Hz. Raw data from an ECG patch was streamed to an android app via Bluetooth during recordings and collected on a central server. Participants also installed a mobile app that collected data from an accelerometer sensor, Bluetooth connectivity, phone and app usage, battery life, and location. An additional app was used for digital questionnaires at specific times during the day. The mobile apps were based on the RADAR platform [14]. During weeks 2 and 4, we asked the participants to wear and test the MHSL-Sleepband v3[1] every night, a wearable device for auditory sleep modulation. We recruited 8 participants aged $\geq$18 years for the RASM pilot study between November 2021 and March 2022 for the 35 days protocol. Of those, three were depressed patients (age 34.3$\pm$14.9, two female) and five non-depressed participants (age 24.4$\pm$0.9, two female).

TABLE I

FEATURES EXTRACTED FOR CLASSIFICATION PER DATA TYPE.

Heart rate	**HRV**	**Physical activity**	**Mobility**
Median	LF	Steps mean	Time at home
Std	HF	Steps std	Radius of gyration
Kurtosis	LFHF	Activity mean	Random entropy
Min	SDNN	Activity std	Max distance home
Max	RMSSD		N. of locations
Entropy	Median NN		Max distance
	SDSD		
	MadNN		
	pNN50		
	Min NN		
	Max NN		

B. Mood Classes

The task for our classification model was to distinguish between four classes, defined as four different emotional states associated with specific regions of the valence (V) - energetic arousal (E) plane, see Fig. 2. This approach is commonly used in psychology research, as it allows for a more nuanced understanding of mood states beyond simple categorization as "happy" or "sad".

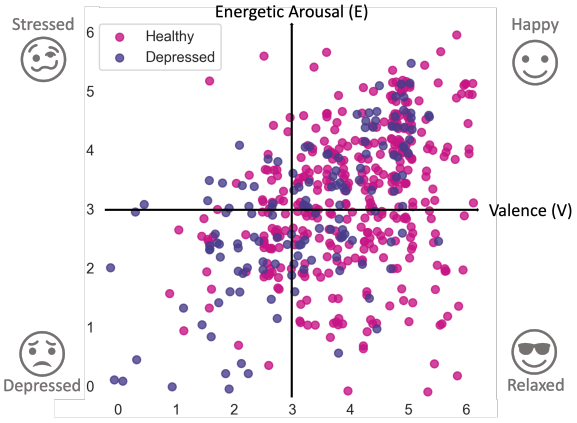

Fig. 2. Energetic Arousal (E) - Valence (V) plane with four class regions. The self-reported labels with values 0-6 are used to identify the corresponding class. The illustration shows the distribution of labels of the healthy participants (pink) and depressed patients (purple).

In the study, we used four questions, two assessing valence (the degree of positivity or negativity) and two assessing energetic arousal from a commonly used and validated 6-item mood questionnaire [15]. We asked participants to complete the mood questionnaire in the morning, after lunch, and in the evening at predefined times (5:00-13:00, 11:00-14:00, and 19:00 - 5:00). The items were rated on a scale from 0 to 6. For analyses, we re-coded the answers so that higher scores indicated higher levels of valence or arousal. We computed four classes of mood states by calculating the mean valence and energetic arousal scores separately. We then plotted the scores on a valence-arousal plane, divided into four quadrants using a threshold of 3. Each quadrant represented a different mood state: "depression" (low V and low E), "stress" (low V and high E), "happiness" (high V and high E), and "relaxation" (high V and low E).

TABLE II

LABEL DISTRIBUTION OF THE HEALTHY PARTICIPANTS (H1-H5) AND DEPRESSED PATIENTS (D1-D3).

Participant	Total	Happy	Stressed	Depressed	Relaxed
H1	**50**	34	10	4	2
H2	**68**	25	5	7	31
H3	**39**	13	7	5	14
H4	**40**	15	6	14	5
H5	**40**	14	5	9	12
D1	**19**	6	3	7	3
D2	**14**	5	0	6	3
D3	**64**	27	3	14	20
Total	**334**	139	39	66	90

C. Data Preprocessing

From the raw data collected, we extracted 28 features from phone sensors and ECG patch; see Table I. In addition, the time of the day (morning, afternoon, and evening) was added as an input feature to give context to the data.

[1]https://www.sleeploop.ch/

We extracted the heart rate (HR) time series from the ECG and calculated simple measures, including the median, standard deviation (std), maximum (max), and minimum (min). We also calculated the entropy. Physical activity metrics, such as the number of steps, were computed with the sensormotion Python package (v.1.1.4) from the 3-axis accelerometer data collected with the ECG Patch.

Heart rate variability (HRV) metrics were extracted using the NeuroKit2 Python package (v. 0.2.2) [16] from a 20 minutes ECG window centered around the time the mood questionnaire was answered. The ECG was first converted to a normal-to-normal (NN) sinus beat interval time series, and then time domain and frequency domain HRV metrics were calculated in the 20-minute segments.

We used GPS phone sensors to gather relative location data from the participants. We then processed this data to extract features from computed trajectories that describe the relative location of the participant, such as the radius of gyration, which measures how far an individual moves around its center of mass (considered the participant's home) [17]. We used the scikit-mobility Python package [18] to extract features from the raw relative location data, sampled every 5 minutes. Data points in the GPS trajectories with speeds higher than 300 km/h from the previous point were considered noisy and removed. We determined the number of visits by counting the total number of points in the trajectory and the number of locations by counting the number of unique points in the trajectory. We also calculated the time spent at home and the maximum distance from home.

To ensure that the extracted features were in a comparable range we applied normalization/scaling. We computed all the features on the data collected on time ranges between mood assessments: morning and afternoon, afternoon and evening, and evening and morning.

D. Classification Pipeline

In this study, we used the Gradient Boosting classifier from scikit-learn (v.1.1.0) library in Python (v.3.9.15) with the default hyperparameter settings of the classifier (learning rate of 0.1, 100 estimators, and a cross-entropy loss). The algorithm calculated the importance of each feature by the Gini importance metric relative to each other.

We also used a naive classifier, trained to consistently predict the majority class present in the training set, as a baseline. We considered this baseline the minimum performance a model should achieve to be considered valuable.

We employed two different cross-validation approaches to evaluate the performance of our proposed model. In the first case, we utilized a leave-one-participant-out (LOPO) strategy, where we excluded the data of a single patient from the training set during each cross-validation iteration (general model).

In the second case, we utilized a leave-part-of-one-participant-out strategy, where only the last week of a participant was excluded from the training set to create a more "personalized" model (Pers_{lw}). These two cross-validation approaches allowed us to evaluate the generalizability and robustness of our model, as well as its ability to handle or adapt to inter- and intra-subject variations in the data. Finally, to have a fair comparison, we repeated the LOPO but tested only on the last week (LOPO_{lw})

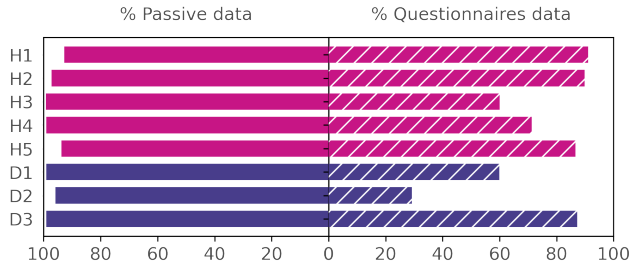

Fig. 3. Percentage of passive and questionnaires data obtained from the healthy participants (H1-H5) and depressed patients (D1-D3) before removing noisy data.

Using the torchmetrics (v.0.11.1) library and macro averaging, we evaluated the accuracy (Acc.), balanced accuracy (B. Acc.), and F1 score. We also computed the Area Under the Precision-Recall Curve (AUPRC) and the Area under the Receiver-Operating Characteristic Curve (AUROC) for the overall multi-class problem.

III. RESULTS

We collected a dataset of passive data and self-reported mood states from 8 participants for up to 35 days, three times per day. Due to technical issues or some participants not completing the full protocol, we ended with 716 samples that had at least either passive data or completed questionnaires. See Fig. 3 for the percentage of passive data and questionnaires obtained by each participant. During the preprocessing step, we removed 104 samples due to missing labels, 9 samples due to missing passive data from both patch and smartphone, 94 missing smartphone data, and 85 missing/noisy patch data. A total of 334 samples were left for analysis of the 8 participants.

Fig. 2 illustrates the distribution of labels across all participants, distinguishing between healthy participants and those diagnosed with depression. Table II shows the distribution of the self-reported mood labels for each participant. The results indicate that the number of self-reported labels was lower among the depressed patients than the non-depressed group.

The results of the classification of the LOPO utilizing passive data are presented in Table III, showing a detailed breakdown of the performance per participant as well as an overall average. The comparison of the different cross-validations is reported in Table III. The baseline model had an average balance accuracy of 0.25 (0.0) and an average AUPRC of 0.26 (0.1), whereas our classifiers had an average balanced accuracy of up to 0.43 (0.12) and AUPRC of 0.41 (0.09). The LOPO and the Personalized models had very similar evaluation metrics.

In Fig. 4, we illustrate the importance of the different features. We computed the importance as the average relative

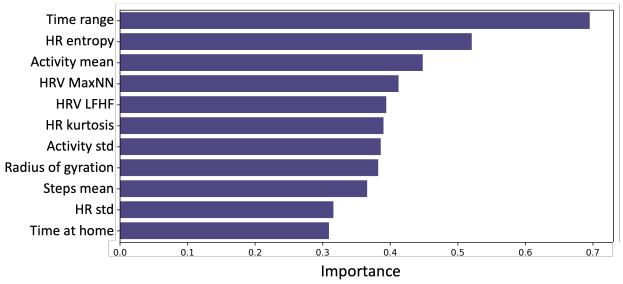

Fig. 4. This figure depicts the average importance of features during the gradient-boosting training process (averaged across all participants). Only features with an importance value above 0.3 are shown.

feature importance (given by the algorithm) between the different test folds for the LOPO. Similar results were obtained for the Pers$_{lw}$ and are not reported.

TABLE III

CLASSIFICATION PERFORMANCE FOR THE LOPO REPORTED FOR EACH FOLD (PARTICIPANT).

Subject	Acc.	B. Acc	F1	AUPRC	AUROC
H1	0.30	0.36	0.20	0.33	0.57
H2	0.51	0.36	0.34	0.35	0.61
H3	0.49	0.35	0.32	0.38	0.62
H4	0.35	0.23	0.18	0.30	0.54
H5	0.45	0.37	0.34	0.49	0.69
D1	0.32	0.29	0.26	0.39	0.55
D2	0.57	0.49	0.43	0.59	0.52
D3	0.31	0.25	0.25	0.31	0.54
Avg.(SD)	0.41(0.11)	0.34(0.08)	0.29(0.08)	0.39(0.10)	0.58(0.06)

TABLE IV

COMPARISON OF CLASSIFICATION METRICS FOR THE BASELINE CLASSIFIER AND FOR THE ONE USING PASSIVE DATA EVALUATED USING LOPO, LOPO$_{lw}$, AND PERS$_{lw}$ CROSS-VALIDATION. RESULTS ARE REPORTED AS AVERAGE (STANDARD DEVIATION) ACROSS FOLDS.

	Baseline	LOPO	LOPO$_{lw}$	Pers$_{lw}$
Acc.	0.4 (0.11)	0.41 (0.11)	**0.43 (0.12)**	0.41 (0.11)
B. Acc.	0.25 (0.00)	**0.34 (0.08)**	**0.34 (0.10)**	**0.34 (0.08)**
F1	0.14 (0.03)	**0.29 (0.08)**	0.28 (0.10)	**0.29 (0.08)**
AUPRC	0.26 (0.03)	0.39 (0.10)	**0.41(0.09)**	0.39 (0.10)
AUROC	0.48 (0.04)	**0.58 (0.06)**	0.53 (0.11)	**0.58 (0.06)**

IV. DISCUSSION

Overall, using models trained on passive data allowed us to classify mood states in our dataset with balanced accuracy and AUPRC scores higher than the baseline model. The small difference between the LOPO$_{lw}$ and Pers$_{lw}$ cross-validation suggests that this approach can help build classifiers adapted to a participant's characteristics while being generalizable to other participants. We did not use models tailored to an individual's data, specifically those trained only on a single participant's data, because some participants did not experience a variety of emotions. These models may not perform well in recognizing emotions that they have not seen before during the training, leading to a lack of generalizability.

The model identified the entropy of HR, average physical activity from the preceding 8 hours, max NN, and NN LFHF ratio at the time of questionnaire completion as the most important features in its analysis. Despite being crucial, the time range of data collection is thought to mainly serve as context, as the distribution of labels is not influenced by the time of day.

While the classifier trained on passive data is not performing as well as we had hoped, it still shows potential and can be improved upon in future work. We have identified several reasons that could contribute to the modest performance of the models. A possible explanation is that the dataset from this pilot study may have a small sample size and a lack of diversity in mood labels among participants, making it challenging for the model to learn and apply to new data. This is because many healthy individuals may not exhibit symptoms of depression or stress. Another reason could be the complexity of the task, as the classification performance is driven by very subjective and biased labels. Unsupervised methods that do not rely on self-reported labels could help overcome this challenge and find unknown patterns for symptom stratification.

It is important to note that collecting data over an extended period, as in this study, only ensures that some emotional states will be captured, as individuals may not consistently or accurately report their moods.

Our data suggest that collecting self-reported labels from depressed patients is more difficult than from healthy individuals due to their difficulty in expressing emotions, and their tendency to withdraw, which may affect the accuracy and reliability of the findings. The added difficulty in collecting labels from depressed patients also poses additional challenges when building machine learning models. As a next step, semi-supervised methods should be explored, as in our study, depressed patients seem to have fewer issues with providing passive data. This also confirms the need to develop unobtrusive tools for monitoring symptoms in the depressed population, especially in the context of clinical trials.

Despite the modest results, our study provides valuable insights and contributes to the growing research on objective mood assessment. Our study provides a benchmark for future research in this area and highlights the challenges and opportunities for improvement in this developing field.

V. CONCLUSIONS

This study aimed at exploring the feasibility of using passive data from a phone and a wearable ECG patch to classify four emotional states in depressed patients and non-depressed participants. The results of this pilot study indicate that it is possible to classify emotional states with slight enhancement over a baseline model. The preliminary results are encouraging and support the potential of this approach to improve the design of objective assessments of interventions that aim at modulating mood and emotional well-being.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to all the participants who took part in this study. Special thanks to

the study assistant C. Schmid who supported us throughout the data collection.

REFERENCES

[1] C. J. Murray and A. D. Lopez, "Evidence-based health policy—lessons from the global burden of disease study," *Science*, vol. 274, no. 5288, pp. 740–743, 1996.

[2] G. Lewis, "Observer bias in the assessment of anxiety and depression," *Social Psychiatry and Psychiatric Epidemiology*, vol. 26, no. 6, pp. 265–272, 1991.

[3] B. W. Dunlop, M. Granros *et al.*, "Recall accuracy for the symptoms of a major depressive episode among clinical trial participants," *Journal of Psychiatric Research*, vol. 116, pp. 178–184, 2019.

[4] E. Dogan, C. Sander *et al.*, "Smartphone-based monitoring of objective and subjective data in affective disorders: where are we and where are we going? Systematic review," *Journal of medical Internet research*, vol. 19, no. 7, p. e7006, 2017.

[5] J. Seppala, I. De Vita *et al.*, "Mobile Phone and Wearable Sensor-Based mHealth Approach for Psychiatric Disorders and Symptoms: Systematic Review and Link to the m-RESIST Project," *JMIR Mental Health*, 2019.

[6] D. A. Rohani, M. Faurholt-Jepsen *et al.*, "Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: systematic review," *JMIR mHealth and uHealth*, vol. 6, no. 8, p. e9691, 2018.

[7] A. Gruenerbl, V. Osmani *et al.*, "Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients," in *Proceedings of the 5th Augmented human international conference*, 2014, pp. 1–8.

[8] A. Grünerbl, A. Muaremi *et al.*, "Smartphone-based recognition of states and state changes in bipolar disorder patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 140–148, 2014.

[9] M. Faurholt-Jepsen, M. Vinberg *et al.*, "Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder," *International journal of methods in psychiatric research*, vol. 25, no. 4, pp. 309–323, 2016.

[10] J. K. Vallance, E. A. Winkler *et al.*, "Associations of objectively-assessed physical activity and sedentary time with depression: NHANES (2005–2006)," *Preventive medicine*, vol. 53, no. 4-5, pp. 284–288, 2011.

[11] G. Valenza, L. Citi *et al.*, "Characterization of depressive states in bipolar patients using wearable textile technology and instantaneous heart rate variability assessment," *IEEE Journal of Biomedical and Heealth Informatics*, vol. 19, no. 1, pp. 263–274, 2014.

[12] K. O. Asare, I. Moshe, *et al.*, "Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status: A longitudinal data analysis," *Pervasive and Mobile Computing*, p. 101621, 2022.

[13] E. Sükei, A. Norbury *et al.*, "Predicting emotional states using behavioral markers derived from passively sensed data: data-driven machine learning approach," *JMIR mHealth and uHealth*, vol. 9, no. 3, p. e24465, 2021.

[14] Y. Ranjan, Z. Rashid *et al.*, "RADAR-base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices," *JMIR mHealth and uHealth*, vol. 7, no. 8, p. e11734, 2019.

[15] P. Wilhelm and D. Schoebi, "Assessing mood in daily life: Structural validity, sensitivity to change, and reliability of a short-scale to measure three basic dimensions of mood." *European Journal of Psychological Assessment*, vol. 23, no. 4, p. 258, 2007.

[16] D. Makowski, T. Pham *et al.*, "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, 2021.

[17] M. C. Gonzalez, C. A. Hidalgo *et al.*, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[18] L. Pappalardo, F. Simini *et al.*, "scikit-mobility: A python library for the analysis, generation and risk assessment of mobility data," *arXiv preprint arXiv:1907.07062*, 2019.