



Politechnika Wroclawska

FIELD OF SCIENCE: Engineering and technology

DISCIPLINE OF SCIENCE: Information and communication technology

DOCTORAL DISSERTATION

Methods for personalized affect recognition from physiological signals in real life

Metody personalizowanego rozpoznawania stanów afektywnych z sygnałów fizjologicznych w życiu codziennym

Bartosz Perz

Supervisor/Supervisors:

prof. dr hab. inż. Przemysław Kazienko

Assistant supervisor:

dr hab. inż. Stanisław Saganowski

Keywords: affect recognition, affective computing, emotion recognition, deep learning, machine learning, personalization, physiological signals, real-life studies

WROCLAW 2024

ACKNOWLEDGEMENTS

After surviving all the struggles I experienced when writing this dissertation, I feel blessed to have met all the good people who helped me during this journey. You opened my eyes to the fact that I am not alone in my struggles and that many people really care about me.

First and foremost, I would like to thank my supervisors, Przemysław Kazienko and Stanisław Saganowski. I am grateful for your guidance and support throughout this journey, as well as your immeasurable patience towards me when I learned from my many mistakes. Without you, writing this dissertation would not be possible.

Secondly, I want to acknowledge everyone involved in the Emognition research group, especially Asia, Dominika, and Maciej. You were awesome collaborators, and I could not ask for better ones. All of the discussions that we had, the problems that we solved, and the time we spent together helped me grow as a researcher and as a person. Also, thank you to programmers, signal processing specialists, students, and others from the Emognition team. Without your work, conducting studies and, thus, my research would have been impossible. Also, thank you to the community of the Department of Artificial Intelligence at Wrocław Tech for creating a wonderful workplace. I appreciate being a part of it.

I have also been fortunate to conduct research with amazing people outside the Emognition team. Working with Nicholas Coles (University of Florida, previously Stanford University) was a great pleasure and a unique opportunity to learn new ways of tackling research problems. I learned a lot from our discussions when designing, conducting, and analyzing the EPiC challenge, and those lessons will stay with me forever. Also, thank you to people from the Signal Analysis and Interpretation Laboratory (SAIL, University of Southern California), especially Prof. Shrikanth Narayanan, for allowing me to visit your lab, learn from you, and conduct the research together.

I would like to also acknowledge my dear family members and friends for all the encouragement I received from you. Having people to complain to and share successes with really helped me survive this journey.

Finally, thank you to Martyna, my dear wife, for your kindness, for supporting me throughout the whole process, caring for me, and helping me stay sane. I can hardly imagine finishing this dissertation without your support.

STRESZCZENIE

W ostatnich latach informatyka afektywna zyskuje na popularności ze względu na swój ogromny potencjał jako dziedziny nauki. Jeżeli udałoby się zrealizować jej założenia, doprowadziłoby to do powstania technologii wnioskujących o ludzkich emocjach, które byłyby prawdziwym przełomem w interakcji człowiek-komputer i potężnym narzędziem do zrozumienia procesów rządzących naszym codziennym życiem. W niniejszej rozprawie zajęliśmy się badaniem metod dla informatyki afektywnej, koncentrując się głównie na personalizowanym rozpoznawaniu stanów afektywnych w życiu codziennym.

W pierwszych pracach skupiliśmy się na przeprowadzeniu dwóch badań literaturowych, t.j., krytycznego omówienia artykułów realizujących rozpoznawanie emocji w sposób aplikowalny w życiu codziennym, oraz przeglądu skupiającego się na procedurach spersonalizowanego wnioskowania o stanach afektywnych. Odkryliśmy dzięki nim konieczność skupienia się w naszej pracy na rozwiązaniach gotowych do użycia w życiu codziennym i na metodach spersonalizowanych. Ponadto, po podsumowaniu głównych różnic między eksperymentami laboratoryjnymi, a badaniami przeprowadzanymi w życiu codziennym, zauważyliśmy nowe wyzwania z którymi badacze muszą się mierzyć poza laboratorium. W niniejszej pracy przedstawiliśmy te wyzwania wraz z rekomendacjami dotyczącymi przyszłych badań nad stanami afektywnymi.

Zajęliśmy się także zgromadzeniem i przygotowaniem dużego zbioru danych z życia codziennego (*LarField*), zawierającego ciągle nagrania wielu sygnałów fizjologicznych i behawioralnych, opatrzonych oznaczeniami stanów emocjonalnych, oraz kontekstem. Jest to jeden z największych nielaboratoryjnych zbiorów danych psychofizjologicznych dotyczących emocji, zawierający nagrania 167 osób, z których każda była monitorowana w ciągu dnia przez jeden miesiąc. Rzeczone dane, wraz z laboratoryjnym zbiorem Emognition, który również zebraliśmy, mogą być wykorzystane do prowadzenia badań nad rozpoznawaniem emocji i innych stanów afektywnych z sygnałów zebranych za pomocą ogólnodostępnych urządzeń noszalnych (ang. *wearables*).

Opracowaliśmy i przetestowaliśmy również nowe metody personalizowanego rozpoznawania stanów afektywnych. Zbadaliśmy, m.in. możliwość wykorzystania modeli spersonalizowanych do rozpoznawania afektu i używania personalizacji grupowej

do przeciwdziałania problemowi zimnego startu (ang. *cold start*). Ponadto opracowaliśmy nową metodę dwuaspektowej personalizacji i zbadaliśmy ją w zadaniu rozpoznawania emocji z sygnałów EKG. Ponieważ przeprowadzane eksperymenty badały różne zjawiska, byliśmy zmuszeni dobierać strategie modelowania, metryki i procedury walidacji osobno do każdego z nich. Nasze badania pokazały przewagę rozwiązań spersonalizowanych nad ogólnymi w dokładności predykcji, zwłaszcza gdy modele były tworzone dla konkretnej osoby, lub gdy wnioskowały one w oparciu o cechy kontekstowe, opisujące osobowość i demografię.

W trakcie prac badawczych mierzyliśmy się także z wieloma wyzwaniami. Największym z nich było radzenie sobie z opóźnieniami w zbieraniu i przetwarzaniu danych z badania LarField. Z ich powodu, personalizacja dwuaspektowa, która została zaprojektowana z myślą o danych zebranych w życiu codziennym, została przetestowana wyłącznie na publicznie dostępnych laboratoryjnych zbiorach danych. Ponadto w badaniach nad przeciwdziałaniem problemowi zimnego startu, ze względu na niską licznosc osobistych próbek każdego z uczestników badania, nie byliśmy w stanie wytrenować w pełni spersonalizowanych modeli. Te wyzwania zainspirowały nas do skomentowania problemów w organizacji badań, nadzorowania ich i zbierania danych psychofizjologicznych dotyczących stanów afektywnych. Szczególną uwagę poświęciliśmy dużym badaniom realizowanym w życiu codziennym, ponieważ nie są one jeszcze tak popularne jak badania laboratoryjne. W związku z tym, wiele problemów z którymi się mierzyliśmy, może nie być znanych innym badaczom.

Niniejsza rozprawa ukazuje istotność wykorzystywania danych zebranych w życiu codziennym i metod personalizowanych w rozpoznawaniu afektu. Pokazuje także potrzebę dalszych badań nad personalizacją, szczególnie nad łączeniem cech indywidualnych z wzorcami obecnymi w całej populacji. W przyszłych eksperymentach planujemy dalszą eksplorację stanów afektywnych oznaczanych przez ludzi, ich związków z fizjologią i sygnałami behawioralnymi, cech personalnych i ogólnych dla populacji, które mogą być wykorzystane do wnioskowania o codziennym życiu, a także nowych metod i strategii dla modelowania stanów afektywnych, w tym modeli bazowych. Większość z tych zagadnień jest eksplorowana przez zespół Emognition, którego jestem członkiem.

ABSTRACT

Affective computing gained much interest in recent years, as its promises, if fulfilled, would lead to creating emotion-aware technologies – a real breakthrough in human-computer interaction, and a powerful tool for understanding processes governing our everyday lives. In this dissertation, we involve ourselves with affective computing, focusing mainly on researching personalized methods for affect recognition in real-life contexts.

Firstly, we performed two literature studies: one critical review of the articles realizing emotion recognition in a manner befitting experiments in everyday life, and another one delving into the procedures for personalized affective computing. They allowed us to discover the necessity of focusing on real-life-ready solutions and personalized methods in our research. Also, as we summarized major differences between classical laboratory experiments and novel in-the-field studies, new challenges introduced by the latter became apparent. We comment on them and give recommendations regarding future endeavors in affective computing.

Other important contributions involve gathering and preparing the *LarField* dataset, a large information-rich dataset collected in everyday life, consisting of continuous recordings of multiple physiological and behavioral signals annotated with emotional states and broad contextual data. It is one of the most extensive datasets on emotion psychophysiology acquired in real life, containing data from 167 subjects, each recorded continuously during their day for one month. This dataset and the in-the-laboratory Emognition dataset that we also produced may be utilized to research affect recognition with off-the-shelf wearable devices.

We also developed and tested new personalized methods for affective computing studies. Among others, we researched the feasibility of using personalized models for affect recognition, utilizing per-group personalization to handle a cold-start problem, and developed a new two-fold personalization method and examined it on emotion recognition from ECG signals. Also, as experiments differed from each other, we had to carefully select modeling strategies, metrics, and validation procedures for each of them. Key findings from our experiments include the superiority of personalized models over general approaches, especially when trained in a subject-specific manner or equipped with features describing personality and demography as a context.

During our research, we experienced several challenges. A major one was handling the delays in data collection and processing from the large field study. Because of that, although two-fold personalization was designed mainly for real-life data, we could only test our methods on laboratory datasets from the literature. Also, in cold-start experiments, we were unable to train fully personalized models due to an insufficient number of per-person samples. These challenges inspired us to comment on the issues of organizing such studies, supervising them, and collecting emotional psychophysiology data. In our lessons learned, we especially focused on extensive outside-the-laboratory studies, as they are still a novelty, and many issues that we faced and shared may not be known to other researchers.

This dissertation emphasizes the importance of utilizing real-life data and respecting subjectivity while designing methods for affect recognition. Moreover, we highlight the need for continued investigations on the balance between the general and individualized modeling approaches. Our future work will focus on further researching patterns in self-reported affective states, their relationships with physiology and behavior, subject-wise and population-wise features that may be utilized for reasoning about people's daily lives, and novel methods and strategies for affect modeling, including foundational models. Most of them have recently been being explored by the Emognition team I am a member of.

RELEVANT SCIENTIFIC PAPERS

- [1] K. Avramidis, D. Kunc, B. Perz, K. Adsul, T. Feng, P. Kazienko, S. Saganowski, and S. Narayanan, “Scaling representation learning from ubiquitous ecg with state-space models,” *IEEE Journal of Biomedical and Health Informatics*, 2024, IF 6.7, MEiN 140 pts.
- [2] N. A. Coles, B. Perz, M. Behnke, J. C. Eichstaedt, S.-H. Kim, T. N. Vu, C. Raman, J. Tejada, G. Zhang, T. Cui, S. Podder, R. Chavda, S. Pandey, A. Upadhyay, J. I. Padilla-Buritica, C. J. Barrera Causil, L. Ji, F. Dollack, K. Kiyokawa, H. Liu, M. Perusquia-Hernandez, H. Uchiyama, X. Wei, H. Cao, Z. Yang, A. Iancarelli, K. McVeigh, Y. Wang, I. M. Berwian, J. C. Chiu, M. Dan-Mircea, E. C. Nook, H. I. Vartiainen, C. Whiting, Y. Won Cho, S.-M. Chow, Z. F. Fisher, Y. Li, X. Xiong, Y. Shen, E. Tagliazucchi, L. Bugnon, R. Ospina, N. M. Bruno, T. A. D’Amelio, F. Zamberlan, L. R. Mercado Diaz, J. O. Pinzon-Arenas, H. F. Posada-Quintero, M. Bilalpur, S. Hinduja, F. Marmolejo-Ramos, S. Canavan, L. Jivnani, and S. Saganowski, “Big team science reveals promises and limitations of machine learning efforts to model the physiological basis of affective experience,” *Nature Human Behaviour*, 2024, In reviews. IF 21.4, MEiN 70 pts.
- [3] J. Komoszyńska, D. Kunc, B. Perz, A. Hebko, P. Kazienko, and S. Saganowski, “Designing and executing a large-scale real-life affective study,” in *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, Biarritz, France, CORE A*, MEiN 200 pts., IEEE, 2024, pp. 505–510.
- [4] D. Kunc, J. Komoszyńska, B. Perz, S. Saganowski, and P. Kazienko, “Emotion system-wearables, physiology, and machine learning for real-life emotion capturing,” in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, MIT Media Lab, Cambridge, MA, USA, IEEE, 2023, 1–3, CORE C, MEiN 20 pts.
- [5] S. Saganowski, B. Perz, A. G. Polak, and P. Kazienko, “Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review,” *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1876–1897, 2023, IF 9.6, MEiN 140 pts. DOI: 10.1109/TAFFC.2022.3176135.
- [6] D. Kunc, J. Komoszyńska, B. Perz, P. Kazienko, and S. Saganowski, “Real-life validation of emotion detection system with wearables,” in *International Work-Conference on the Interplay Between Natural and Artificial Computation (IWINAC)*, Puerto de la Cruz, Tenerife, Spain, Springer, 2022, 45–54, CORE C, MEiN 20 pts.
- [7] B. Perz, “Personalization of emotion recognition for everyday life using physiological signals from wearables,” in *2022 10th International Confer-*

ence on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), Nara, Japan, IEEE, 2022, 1–5, CORE C, MEiN 20 pts.

- [8] S. Saganowski, J. Komoszyńska, M. Behnke, B. Perz, D. Kunc, B. Klich, Ł. D. Kaczmarek, and P. Kazienko, “Emognition dataset: Emotion recognition with self-reports, facial expressions, and physiology using wearables,” *Scientific data*, vol. 9, no. 1, p. 158, 2022, IF 9.8, MEiN 140 pts. DOI: [10.1038/s41597-022-01262-0](https://doi.org/10.1038/s41597-022-01262-0).
- [9] S. Saganowski, D. Kunc, B. Perz, J. Komoszyńska, M. Behnke, and P. Kazienko, “The cold start problem and per-group personalization in real-life emotion recognition with wearables,” in *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, WristSense 2022 - The Eighth Workshop on Sensing Systems and Applications Using Wrist Worn Smart Devices, CORE A*, MEiN 200 pts., Best Paper Award, IEEE, 2022, pp. 812–817. DOI: [10.1109/PerComWorkshops53856.2022.9767233](https://doi.org/10.1109/PerComWorkshops53856.2022.9767233).
- [10] 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 *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, online, IEEE, 2021, 1–3, CORE C, MEiN 20 pts.

TABLE OF CONTENTS

Acknowledgements	iii
Streszczenie	v
Abstract	vii
Relevant scientific papers	ix
Table of Contents	x
List of Illustrations	xiii
List of Tables	xvi
Nomenclature	xxiii
Chapter I: Introduction	1
1.1 Contributions and achievements	3
1.2 Dissertation outline	6
1.3 Other remarks	6
Chapter II: Affective computing and physiology	8
2.1 Emotions	9
2.2 Physiology	15
Chapter III: Critical literature review	19
3.1 Methodology	20
3.2 Emotion recognition study design	22
3.3 Study participants	27
3.4 Temporal context affecting emotion	28
3.5 Collecting physiological signals	30
3.6 Emotional models and adjustments	32
3.7 Emotion labeling	34
3.8 From theoretical model to reasoning task	37
3.9 <i>Learning case</i> definition	39
3.10 Machine learning models	40
3.11 Model validation	42
3.12 Discussion and identified challenges	46
Chapter IV: Personalized emotion recognition review	55
4.1 Personalization strategies	57
4.2 Signals for personalized ER	63
4.3 Validation in personalized ER	65
4.4 Discussion	68
Chapter V: In-the-laboratory data collection	73
5.1 Experimental procedure	74
5.2 Results	76
5.3 Discussion	78
Chapter VI: Cold start and group personalization	83
6.1 Materials and methods	83

6.2 Results	87
6.3 Discussion	91
Chapter VII: Emotions in the wild	92
7.1 Designing a large real-life psychophysiology study	92
7.2 Conducting a study	97
7.3 Results	102
7.4 Discussion	112
Chapter VIII: Searching for physiological markers of emotion	117
8.1 Competition as research method	118
8.2 Materials and methods	119
8.3 Results	125
8.4 Discussion	129
Chapter IX: Personalized data processing	132
9.1 Materials and methods	132
9.2 Results	139
9.3 Discussion	157
Chapter X: Conclusions	159
10.1 Affect recognition literature	159
10.2 Collecting data for emotion research	160
10.3 Personalization for real-life affect recognition	161
10.4 Limitations of this work and affective computing research	162
10.5 Summary and future work	163
Bibliography	166
Appendix A: Critical literature review	188
Appendix B: In-the-laboratory data collection	196
Appendix C: Emotions in the wild	199
C.1 Questionnaires utilized for experiments	199
C.2 Results	205
Appendix D: EPiC competition submissions	228
Appendix E: Personalized emotion recognition	236
E.1 Proof for z-score equality	236
E.2 Personalized processing	237

LIST OF ILLUSTRATIONS

<i>Number</i>	<i>Page</i>
2.1 Plutchik’s wheel of emotions.	11
2.2 Circumplex model of emotions.	12
2.3 Pleasure-Arousal-Dominance (PAD) emotion model.	12
2.4 Interrelationships between physiological systems and biosignals. . . .	15
3.1 Number of papers in relation to the year of publication and study environment.	22
3.2 Common and unique research stages (study design) for emotion recognition in the lab and in the field.	23
3.3 Emotion recognition scenarios identified in SLR.	26
3.4 Possible ways of labeling physiological signals with emotions in the lab studies.	36
3.5 Possible ways of triggering self-assessment in the field studies.	36
3.6 Transition of original psychological emotional models into machine learning used in experiments and conversion of 2-dimensional emo- tional <i>arousal-valence</i> space to machine learning problem.	38
3.7 Selected validation methods used in emotion recognition.	45
5.1 Examples of data available in the Emognition dataset.	81
5.2 Distribution of self-reported emotions between and within conditions.	82
6.1 Scenarios considered in cold start and group personalization experi- ments.	84
6.2 Mean F1 scores for AdaBoost classifier.	88
7.1 Timeline of the LarField phases: designing and main study.	95
7.2 Devices used in the LarField study.	95
7.3 Distributions of Big Five personality model scores in LarField dataset.	104
7.4 Results of morning questionnaire prediction on Larfield dataset. . . .	106
7.5 Results of evening questionnaire prediction on Larfield dataset. . . .	107
7.6 Results of emotion questionnaires regression on Larfield dataset in daily setup.	109
7.7 Results of emotion questionnaires classification on Larfield dataset in daily setup.	110

7.8	Results of emotion questionnaires prediction on Larfield dataset in momentary setup.	111
8.1	Validation scenarios employed in EPiC competition.	121
8.2	Mean absolute prediction error for predictions submitted to EPIC competition.	127
8.3	Example of real and simulated random electrocardiography signal and mean absolute prediction errors for additional testing.	128
9.1	Distribution of annotations in a subset of AMIGOS dataset used in experiments.	133
9.2	Distribution of annotations in a subset of ASCERTAIN dataset used in experiments.	134
9.3	Distribution of annotations in a subset of CASE dataset used in experiments.	136
9.4	Distribution of annotations in a subset of DREAMER dataset used in experiments.	137
9.5	Self-reported arousal and valence of selected subjects from CASE dataset before and after adjusting.	141
9.6	Comparison of model training approaches with baselines on classification task in group setup.	142
9.7	Comparison of model training approaches with baselines on regression task in group setup.	143
9.8	Comparison of processing methods on AMIGOS dataset in group training setup.	144
9.9	Comparison of processing methods on ASCERTAIN dataset in group training setup.	145
9.10	Comparison of processing methods on CASE dataset in group training setup.	146
9.11	Comparison of processing methods on DREAMER dataset in group training setup.	147
9.12	Comparison of model training approaches with baselines on classification task in subject-specific setup.	148
9.13	Comparison of model training approaches with baselines on classification task in subject-specific setup.	149
9.14	Comparison of processing methods on AMIGOS dataset in subject-specific training setup.	150

9.15	Comparison of processing methods on ASCERTAIN dataset in subject-specific training setup.	151
9.16	Comparison of processing methods on CASE dataset in subject-specific training setup.	152
9.17	Comparison of processing methods on DREAMER dataset in subject-specific training setup.	153
9.18	Comparison of experimental designs with baselines on classification tasks.	154
9.19	Comparison of experimental designs with baselines on regression tasks.	155
B.1	The original (Polish) version of self-reports used in the Emognition study.	196
C.1	Personality questionnaire used in the Emognition study.	200
C.2	Morning questionnaire used in the Emognition study.	203
C.3	Morning questionnaire used in the Emognition study.	203
C.4	Emotion questionnaire used in the Emognition study.	204

LIST OF TABLES

<i>Number</i>	<i>Page</i>
2.1 Stages of affective computing development.	8
2.2 Categories of emotional cues for affective research.	13
3.1 A number of articles excluded from SLR with respect to particular inclusion or exclusion criteria.	21
3.2 Main differences between laboratory and real-life emotion recognition studies.	24
3.3 Participants' metadata surveyed in studies.	27
3.4 Context surveyed or considered in reviewed studies.	30
3.5 Physiological signals used for emotion recognition.	31
3.6 The most popular physiology-measuring devices in SLR.	31
3.7 Trigger times, types of self-assessment, and emotional models utilized in studies.	33
3.8 Methods of emotion labeling (ground truth).	36
3.9 Possible ways of collecting self-assessment.	37
3.10 Approaches, methods, and measures used at particular machine learning stage.	53
3.11 Validation methods.	54
4.1 Personalization approaches in reviewed studies.	56
4.2 Machine learning approaches in reviewed studies.	62
4.3 Affective states researched in reviewed studies.	63
4.4 Signals and devices used for affect recognition in reviewed studies.	64
4.5 Approaches to validation in reviewed studies.	66
5.1 Measures available in the Emognition dataset.	75
5.2 Signal-to-noise ratios (SNRs) statistics computed for raw physiological signals recorded during the study.	77
6.1 Distribution of data used in the research.	84
6.2 Features extracted from the physiological signals.	86
6.3 Results for investigated scenarios - classifiers using only cardiac data as input.	89
6.4 Results for investigated scenarios - classifiers using cardiac data and acceleration (ACC) as input.	90
7.1 Signals and measures collected during the LarField study.	94
7.2 Main aspects validated in pilot studies.	99

7.3	Results for emotion measure regression from personality trait scores.	105
7.4	Selected challenges encountered during the design and execution of the study, along with the suggested solutions.	113
A.1	Emotional models, ground truth, and machine learning (ML) problems.	189
A.2	Main differences in emotion recognition between lab study and field study.	192
A.3	Machine learning-related techniques and measures applied in SLR studies.	193
B.1	Results of Repeated Measures Analysis of Variance for Differences Between Conditions in Self-reported Emotions.	197
B.2	Results of Repeated Measures Analysis of Variance for Differences Within Conditions in Self-reported Emotions.	197
B.3	<i>P</i> -values from repeated measures analysis of variance (rANOVA) for differences between conditions	198
B.4	<i>P</i> -values from repeated measures analysis of variance (rANOVA) for differences within conditions	198
C.1	Results for regression of next day morning sleep quality questionnaire.	205
C.2	Results for regression of next day morning rest questionnaire.	206
C.3	Results for regression of next day morning stress questionnaire.	206
C.4	Results for regression of next day morning composure questionnaire.	207
C.5	Results for classification of next day morning sleep quality questionnaire.	207
C.6	Results for classification of next day morning rest questionnaire.	208
C.7	Results for classification of next day morning stress questionnaire.	208
C.8	Results for classification of next day morning composure questionnaire.	209
C.9	Results for regression of same day evening health questionnaire.	209
C.10	Results for regression of same day evening mood questionnaire.	210
C.11	Results for regression of same day evening overwhelm questionnaire.	210
C.12	Results for regression of same day evening unpredictability questionnaire.	211
C.13	Results for classification of same day evening health questionnaire.	211
C.14	Results for classification of same day evening mood questionnaire.	212
C.15	Results for classification of same day evening overwhelm questionnaire.	212
C.16	Results for classification of same day evening unpredictability questionnaire.	213
C.17	Results for regression of daily emotions morning valence questionnaire.	213

C.18	Results for regression of daily emotions morning arousal questionnaire.	214
C.19	Results for regression of daily emotions afternoon valence questionnaire.	214
C.20	Results for regression of daily emotions afternoon arousal questionnaire.	215
C.21	Results for regression of daily emotions evening valence questionnaire.	215
C.22	Results for regression of daily emotions evening arousal questionnaire.	216
C.23	Results for classification of daily emotions morning valence questionnaire.	216
C.24	Results for classification of daily emotions morning arousal questionnaire.	217
C.25	Results for classification of daily emotions afternoon valence questionnaire.	217
C.26	Results for classification of daily emotions afternoon arousal questionnaire.	218
C.27	Results for classification of daily emotions evening valence questionnaire.	218
C.28	Results for classification of daily emotions evening arousal questionnaire.	219
C.29	Results for regression of momentary emotions valence questionnaire.	219
C.30	Results for regression of momentary emotions arousal questionnaire.	220
C.31	Results for classification of momentary emotions intense emotions questionnaire.	220
C.32	Results for classification of momentary emotions valence questionnaire.	221
C.33	Results for classification of momentary emotions arousal questionnaire.	221
C.34	Pairwise comparisons of modeling strategies for regression of next day morning questionnaires.	223
C.35	Pairwise comparisons of modeling strategies for classification of next day morning questionnaires.	223
C.36	Pairwise comparisons of modeling strategies for regression of same day evening questionnaires.	224
C.37	Pairwise comparisons of modeling strategies for classification of same day evening questionnaires.	224
C.38	Pairwise comparisons of modeling strategies for regression of daily emotions questionnaires.	225
C.39	Pairwise comparisons of modeling strategies for classification of daily emotions questionnaires.	226

C.40	Pairwise comparisons of modeling strategies for regression of momentary emotions questionnaires.	226
C.41	Pairwise comparisons of modeling strategies for classification of momentary emotions questionnaires.	227
D.1	Details of submissions to Emotion Physiology and Experience Collaboration (EPiC) challenge.	228
D.2	Competition results in across-time validation scenario.	231
D.3	Competition results in across-subject validation scenario.	232
D.4	Competition results in across-emotion validation scenario.	233
D.5	Competition results in across-induction validation scenario.	234
D.6	Results for the three teams selected to partake in additional testing.	235
E.1	Classification results in subject-dependent (group) experimental setup	238
E.2	Regression results in subject-dependent (group) experimental setup	239
E.3	Results of Friedman’s test between processing methods in subject-dependent (group) experimental design.	240
E.4	Comparison of processing methods in subject-dependent (group) experimental design. AMIGOS dataset, classification, majority baseline.	241
E.5	Comparison of processing methods in subject-dependent (group) experimental design. AMIGOS dataset, regression, average baseline.	242
E.6	Results of Conover’s post-hoc test between processing methods in subject-dependent (group) experimental design. AMIGOS dataset, arousal classification.	243
E.7	Comparison of processing methods in subject-dependent (group) experimental design. ASCERTAIN dataset, classification, majority baseline.	244
E.8	Comparison of processing methods in subject-dependent (group) experimental design. ASCERTAIN dataset, regression, average baseline.	245
E.9	Results of Conover’s post-hoc test between processing methods in subject-dependent (group) experimental design. ASCERTAIN dataset, arousal classification.	246
E.10	Results of Conover’s post-hoc test between processing methods in subject-dependent (group) experimental design. ASCERTAIN dataset, valence classification.	247
E.11	Results of Conover’s post-hoc test between processing methods in subject-dependent experimental design. ASCERTAIN dataset, valence regression.	248

E.12	Comparison of processing methods in subject-dependent (group) experimental design. CASE dataset, classification, majority baseline.	249
E.13	Comparison of processing methods in subject-dependent (group) experimental design. CASE dataset, regression, average baseline.	250
E.14	Results of Conover's post-hoc test between processing methods in subject-dependent experimental design. CASE dataset, arousal regression.	251
E.15	Comparison of processing methods in subject-dependent (group) experimental design. DREAMER dataset, classification, majority baseline.	252
E.16	Comparison of processing methods in subject-dependent (group) experimental design. DREAMER dataset, regression, average baseline.	253
E.17	Classification results in subject-dependent (subject) experimental setup.	255
E.18	Regression results in subject-dependent (subject) experimental setup.	256
E.19	Results of Friedman test between processing methods in subject-dependent (subject) experimental design.	257
E.20	Comparison of processing methods in subject-dependent (subject) experimental design. AMIGOS dataset, classification, majority baseline.	258
E.21	Comparison of processing methods in subject-dependent (subject) experimental design. AMIGOS dataset, regression, average baseline.	259
E.22	Results of Conover's post-hoc test between processing methods in subject-dependent (subject) experimental design. AMIGOS dataset, arousal regression.	260
E.23	Results of Conover's post-hoc test between processing methods in subject-dependent (subject) experimental design. AMIGOS dataset, valence regression.	261
E.24	Comparison of processing methods in subject-dependent (subject) experimental design. ASCERTAIN dataset, classification, majority baseline.	262
E.25	Comparison of processing methods in subject-dependent (subject) experimental design. ASCERTAIN dataset, regression, average baseline.	263
E.26	Results of Conover's post-hoc test between processing methods in subject-dependent (subject) experimental design. ASCERTAIN dataset, arousal classification.	264

E.27	Results of Conover's post-hoc test between processing methods in subject-dependent (subject) experimental design. ASCERTAIN dataset, valence classification.	265
E.28	Results of Conover's post-hoc test between processing methods in subject-dependent (subject) experimental design. ASCERTAIN dataset, arousal regression.	266
E.29	Comparison of processing methods in subject-dependent (subject) experimental design. CASE dataset, classification, majority baseline.	267
E.30	Comparison of processing methods in subject-dependent (subject) experimental design. CASE dataset, regression, average baseline. . .	268
E.31	Results of Conover's post-hoc test between processing methods in subject-dependent (subject) experimental design. CASE dataset, arousal classification.	269
E.32	Results of Conover's post-hoc test between processing methods in subject-dependent (subject) experimental design. CASE dataset, arousal regression.	270
E.33	Comparison of processing methods in subject-dependent (subject) experimental design. DREAMER dataset, classification, majority baseline.	271
E.34	Comparison of processing methods in subject-dependent (subject) experimental design. DREAMER dataset, regression, average baseline.	272
E.35	Results of Conover's post-hoc test between processing methods in subject-dependent (subject) experimental design. DREAMER dataset, arousal classification.	273
E.36	Results of Conover's post-hoc test between processing methods in subject-dependent (subject) experimental design. DREAMER dataset, arousal regression.	274
E.37	Comparison of results in experimental setups (designs) with respective baselines. Arousal classification, majority baseline.	276
E.38	Comparison of results in experimental setups (designs) with respective baselines. Valence classification, majority baseline.	277
E.39	Comparison of results in experimental setups (designs) with respective baselines. Arousal regression, average baseline.	278
E.40	Comparison of results in experimental setups (designs) with respective baselines. Valence regression, average baseline.	279
E.41	Results of Friedman test between experimental designs.	280

E.42	Results of Conover's post-hoc test between experimental designs. AMIGOS dataset, Arousal classification.	280
E.43	Results of Conover's post-hoc test between experimental designs. AMIGOS dataset, Valence classification.	281
E.44	Results of Conover's post-hoc test between experimental designs. AMIGOS dataset, Valence regression.	281
E.45	Results of Conover's post-hoc test between experimental designs. ASCERTAIN dataset, Arousal classification.	281
E.46	Results of Conover's post-hoc test between experimental designs. ASCERTAIN dataset, Valence classification.	282
E.47	Results of Conover's post-hoc test between experimental designs. ASCERTAIN dataset, Arousal regression.	282
E.48	Results of Conover's post-hoc test between experimental designs. ASCERTAIN dataset, Valence regression.	282
E.49	Results of Conover's post-hoc test between experimental designs. CASE dataset, Arousal classification.	283
E.50	Results of Conover's post-hoc test between experimental designs. CASE dataset, Valence classification.	283
E.51	Results of Conover's post-hoc test between experimental designs. CASE dataset, Arousal regression.	283
E.52	Results of Conover's post-hoc test between experimental designs. CASE dataset, Valence regression.	284
E.53	Results of Conover's post-hoc test between experimental designs. DREAMER dataset, Arousal classification.	284
E.54	Results of Conover's post-hoc test between experimental designs. DREAMER dataset, Valence classification.	284
E.55	Results of Conover's post-hoc test between experimental designs. DREAMER dataset, Valence regression.	285

NOMENCLATURE

- Accuracy (general term).** The degree to which predictions or results match the correct values.
- Accuracy (metric).** A fraction of correctly classified instances (true positives + true negatives) to all predictions.
- Adjacent sliding windows.** Signal windows shifted by the whole window length or more between. As a result, obtained parts do not contain common samples.
- Arousal.** Emotional dimension describing how energized or soporific a person feels.
- AUC.** Area under curve, usually calculated for Receiver Operating Characteristic (ROC) curve.
- CCC.** Concordance correlation coefficient.
- CNN.** Convolutional Neural Network.
- DT.** Decision Tree.
- Emotional stimulus.** Any event (or object) that elicits emotional experience.
- F1-measure (score).** A harmonic mean of precision and recall.
- KNN.** K-nearest neighbors.
- LO(k)SO.** Leave One (or k) Subjects Out cross validation.
- LSTM.** Long short-term memory network.
- Macro averaging (-macro).** Computing arithmetic mean across classes, i.e., first computing performance metric values per class (for instance, high class and low class precision), and afterwards averaging the class-wise scores..
- MAE.** Mean absolute error.
- MLP.** Multilayer Perceptron.
- MSE.** Mean squared error.
- Overfitting.** A situation where trained model does not capture relationships between input and output data, but learns to mimick the exact input-output pairs seen during training.
- Overlapping sliding windows.** Signal windows shifted by fewer samples than window length. As a result, obtained parts contain common samples.

Precision. A fraction of correctly classified positive class instances (true positives) to all instances classified as positive class (true positives + false positives).

Recall. A fraction of correctly classified positive class instances (true positives) to all of positive class instances in the dataset (true positives + false negatives).

RMSE. Square root of mean squared error.

Sliding windows. A method where an operation is executed within a predefined window (or range) that is shifted across the input data. Often used to perform computations over input signals, or to divide signal into parts for later processing..

SMS. Short Message Service, a service for text messaging implemented in most telephones.

SVM. Support Vector Machine.

Timestamp. A number coding a specific moment in time, usually representing the amount of seconds elapsed since midnight on 1 January 1970 UTC (the Unix Epoch).

Valence. Emotional dimension describing how positive or negative a person feels (or how pleasant or unpleasant a stimulus is).

Chapter 1

INTRODUCTION

In the span of our lives, we people experience countless occurrences of phenomenological states that we call “emotions” [1]. Those states drive most of our decisions, not only those intuitive but also the seemingly logical ones, and also impact perception and cognitive functions [2–4]. It is thus not surprising that technological companies pursue, among others, emotion recognition technologies and that the emotion AI industry was recently valued at over 20 billion USD [5], with some sources predicting its further growth [6].

To realize emotion recognition, several obstacles have to be overcome first. One of them regards finding patterns that allow discerning between different emotional states. Although it is well-known that affective experience impacts various systems within the human body, the existence of universal links between experienced emotions and nonverbal cues, such as behavior or physiology, is still subject to debate [7–10]. Different authors point out personal differences in both physiology and emotion perception, along with the impact of context on experience [11–14]. The lack of consensus is best expressed by the fact that most researchers still work on models for the general population while others focus on creating personalized solutions.

It leads to another set of issues regarding the relatively low amounts of personal data that can be collected, as people tend to become tired if asked about their emotions too often [15]. Not only that, but not having any data on a person for whom one wants to predict emotions does not allow adjusting and improving models, and it is called the cold start problem. This issue does not affect general models, but in personalization research, one can also find ways to mitigate it by, e.g., modeling patterns within groups of similar people or utilizing procedures for low-data regimes, such as creating unsophisticated models or using transfer learning.

Emotion recognition systems may benefit the general population in a variety of ways, by, e.g., improving human-computer interaction, helping teachers in the classroom, helping diagnose patients and provide custom treatments, assisting in decision making processes, or improving people’s wellbeing and longevity [16]. It could also improve computational reasoning, similar to that of humans, by learning to opti-

mize repetitive tasks or by learning how to make decisions when problems cannot be fully explored [10]. However, there is also an ongoing discussion about harms that unwarranted emotion recognition or inaccurate predictions may cause [17, 18].

Possibly helpful and destructive at the same time, algorithms for emotion recognition should not only be precise but also created with respect to people's privacy in mind. While multimodal systems would probably work best, as people express emotions in various ways, modalities differ in terms of their (1) susceptibility to noise and (2) privacy. For example, body gestures and facial expressions can be suppressed or changed [19], and many researchers perceive them as impacted by context, such as culture or specific situations [20–23]. Moreover, facial, behavioral, and vocal expressions can be tracked without a person's consent through cameras and microphones, which are becoming increasingly ubiquitous. Unlike those, recording physiological data, such as skin temperature or heart rate, typically requires direct contact with the person's body and has to be realized, e.g., by using wearable sensors. Picard [10] raised similar concerns in her book and pointed out that general and ubiquitous methods may not be the holy grail of emotion recognition, as some people may prefer only their trusted devices and closest ones to access their emotional information.

Although those issues cannot be ignored, such ethical considerations are not in the scope of this thesis, and therefore, we do not delve deeply into them. However, having all of the concerns in mind, in this dissertation, we mostly focus on methods for affect recognition that (1) may be applied in everyday life and (2) utilize signals that are difficult to measure without a person's knowledge. While in review sections, the reader can find information about a vast range of methods, in our own research, we focus solely on physiological signals and their application for emotion recognition. Additionally, because during our research, we found many issues with commensurability, we also explored different validation and training procedures suitable for emotion recognition studies. Specifically, this thesis tries to answer the following fundamental questions:

- Is there a necessity for affective studies to be transferred from controlled laboratory conditions to real life? What are the main differences between the laboratory and field studies, and what new challenges does the latter introduce? (Chap. 3)

- Can everyday life affective states be recognized from physiological signals? (Chaps. 7)
- Does a personalized approach to affect recognition improve the quality of inference of emotional states from physiological signals over the commonly utilized generalized methods and baselines? (Chaps. 7, 6, and 9)

1.1 Contributions and achievements

While a substantial amount of work still needs to be done in affective computing and emotion recognition fields, in this dissertation, we provide the following contributions and achievements:

1. To explore methods and issues in emotion recognition research, we performed a critical systematic review, focusing on studies performed in or applicable to real life ([16], Chap. 3).
2. Additionally, we reviewed the literature regarding personalized affect recognition, starting from personalized context impacting emotional experience ([16] - Appendix, Secs. 3.3 and 3.4). Finally, we structured the topic, identified research components that may be subject to personalization, and critically revised possible approaches (Chap. 4).
3. We also identified key between-people differences that have an impact on the quality of affect recognition methods, and underlined the importance of developing personalized solutions ([16], Secs. 3.3 and 3.4). These dissimilarities can be seen, among others, in people's (1) perception and understanding of affective states, (2) physiological reactions to stimuli, (3) engagement and attitude towards the data collection process, (4) personalities, which impact emotional reactions, and (5) lifestyle and everyday habits.
4. We collected a laboratory dataset utilizing off-the-shelf devices suitable for use in real-life studies ([24], Chap. 5). In particular, a new lessons learned section (Sec. 5.3.2) was added in this thesis¹.
5. The above dataset served as preliminary work before running a large real-life psychophysiology study (LarField), for which we provide descriptions and

¹Although our own experiments did not yield satisfactory results, other researchers found the dataset valuable, and at the time of writing (August 29, 2024) it was cited 50 times according to google scholar and 33 times according to Scopus (with self-citations excluded).

recommendations based on the lessons learned ([25], Chap. 7). Data from this study (LarField dataset) was collected using using the Emognition system that we personally contributed to ([15, 26, 27], Sec. 7.1.2).

6. On the LarField dataset, we performed initial experiments focused on studying the feasibility of including personal context when modeling real-life affective states. Utilized methods, albeit simple, allowed us to draw some conclusions about personalized approaches to recognition of daily affective states and momentary emotions, which we present together with identified limitations (Secs. 7.3 and 7.4.2).
7. Moreover, using data collected during in-the-field pilot studies preceding the collection of the LarField dataset, we researched different strategies of training models for emotion detection and the usefulness of group-personalization as a means of addressing the cold-start problem in emotion recognition studies ([28], Chap. 6).
8. For the above research on group-personalization and the conference paper describing it [28], we were granted the *Best Paper Award at WristSense 2022 - The Eighth Workshop on Sensing Systems and Applications Using Wrist Worn Smart Devices*, co-located with *2022 IEEE International Conference on Pervasive Computing and Communications* (CORE A*, MEIN 200 pts.).
9. We researched the properties of the WildECG, a pre-trained state-space (S4) model for representation learning from ECG signals, in collaboration with the University of Southern California [29]. We used this pre-trained model in our research of two-fold personalization to generate signal representations (Sec. 9.1.4).
10. We researched different methods of personalization for emotion recognition from ECG, using new two-fold personalization strategies and the WildECG model. They focused mainly on personalized standardization of input signals and collected emotion measures. We present the results of our research and provide conclusions about the impact that different normalization methods have on emotion recognition quality (Chap. 9).
11. Additionally, on the same datasets, we analyzed the effect of personalized training on results and compared two personalized experimental designs with non-personalized models (Chap. 9).

12. We co-lead a *Big Team Science* effort in the form of a competition and a workshop at the 2023 Affective Computing and Intelligent Interaction Conference (ACII 2023, MIT, Boston, USA)^{2,3,4}, aimed to evaluate the possibilities and restrictions that utilizing machine learning methods to model theorized links between PNS activity and emotion self-reports with researchers from Stanford University (California) and Adam Mickiewicz University (Poznań, Poland) ([32], Chap. 8).
13. After the above challenge concluded, we ranked the submissions of competition participants and investigated the best-performing methods, by reviewing them and rerunning the submitted code to assess their replicability ([32], Chap. 8).
14. Additionally, we investigated the challenge’s results and drew observations and conclusions regarding the comparability of different machine learning algorithms and methods, and further, we provide comments regarding constraints on generalizability imposed by a chosen approach. This work was done with researchers from, among others: Adam Mickiewicz University (Poland), Chonnam National University (South Korea), Delft University of Technology (Netherlands), Federal University of Sergipe (Brazil), FPT University (Vietnam), Indian Institute of Science Education and Research (India), Indian Institute of Technology (India), Instituto Tecnológico Metropolitano (Colombia), Montana State University (MA, US), Nara Institute of Science and Technology (Japan), New York Institute of Technology (NY, US), Northeastern University (MA, US), Princeton University (NJ, US), The Pennsylvania State University (PA, US), Stanford University (CA, US), Tilburg University (Netherlands), Universidad Adolfo Ibáñez (Chile), Universidad Nacional del Litoral (Argentina), Universidade Federal da Bahia (Brazil), University of Buenos Aires (Argentina), University of Connecticut (CT, US), University of Pittsburgh (PA, US), University of South Australia Online (Australia), University of South Florida (FL, US), and one independent researcher ([32], Chap. 8).
15. In total, research that we published was cited⁵: 156 times, with an h-index of

²<https://epic-collab.github.io/acii/>

³During the event, we personally delivered a speech describing the challenge structure [30], which we later presented in an extended version during 2023 Big Team Science Conference (BTSCON) [31]

⁴Also, during the workshop we had an honor to host Lisa Feldman Barrett as a keynote speaker

⁵Search performed on September 26, 2024.

5, according to Google Scholar; or 103 times (72 times without autocitations), with an h-index of 2, according to Scopus.

1.2 Dissertation outline

Chapter 2 gives the core background and definitions necessary to understand the rest of this dissertation and the affective computing literature in general.

Chapter 3 contains the literature research focused on emotion recognition for everyday life.

Chapter 4 contains the literature research focused on personalized methods for affect recognition.

Chapter 5 describes procedures, results, and lessons learned from collecting laboratory emotion psychophysiology dataset.

Chapter 6 describes our research on per-group personalization and its use to mitigate the cold start problem in emotion recognition studies.

Chapter 7 describes procedures, results, and lessons learned from collecting real-life emotion psychophysiology dataset.

Chapter 8 provides details of the Emotion Physiology and Experience Collaboration (EPiC) challenge, together with obtained results and conclusions regarding Big Team Science in affective computing, commensurability and generalizability challenges, and observation regarding the multiplicative nature of constraints on generalizability.

Chapter 9 contains details of conducted research on two-fold personalization for emotion recognition.

Chapter 10 summarizes the thesis and comments on obtained results. We also discuss future work that could be conducted to improve upon the presented methods.

1.3 Other remarks

This dissertation was prepared using a Caltech thesis template [33].

Figures and tables indexed with numbers only are located in the main body of this dissertation. The ones indexed with a letter first, are located in the respective appendixes (e.g., Tab A.1 is located in Appendix A).

Contents of this dissertation should be treated as our own work (mine with support from my supervisors) unless otherwise noted.

The dissertation's text was written with the assistance of programs employing artificial intelligence, which supported us in spellchecking, correcting language mistakes, and improving readability while preserving the original sense of sentences.

Chapter 2

AFFECTIVE COMPUTING AND PHYSIOLOGY

When talking about affective computing, many researchers cite Picard [10], stating that affective computing "relates to, arises, or deliberately influences emotions." In practice, affect is used as an umbrella term for many different feelings and internal states, e.g., mood (long-lasting feelings) [34, 35], or any basic feelings [36, 37]. Additionally, affect is heavily influenced by other internal states of a person, e.g., stress, well-being, and health [38]. Consequently, affective computing is a term and a research field encompassing different calculations concerning people's subjective experiences, aimed at introducing affective (or emotional) intelligence to computers, along with abilities to recognize or express affect (Tab. 2.1). However, despite the great effort researchers have poured into the field for over two decades, we have yet to reach any significant milestone (expressing or perceiving emotions in a human-like fashion).

Table 2.1: Stages of affective computing development, relative to computers' affective abilities (based on [10]).

Computer abilities regarding affect	Cannot express	Can express
Cannot perceive	I	II
Can perceive	III	IV

Schmidt et al. [12] described affect recognition as an interdisciplinary research field, utilizing knowledge from psychology, neuroscience, machine learning, and signal processing to find patterns connecting affective states with its indicators. In this work, we focus mainly on a subfield of affective computing, i.e., recognition of short-lasting intense affective states called emotions¹ [8]. Since the goal of affect recognition is to identify a person's affective state (e.g., emotion, mood, stress) based on some observable indicators [12], emotion recognition restricts this scope to recognizing only emotional states. Building on the view presented by Schmidt et al. [12], who described affect recognition as an interdisciplinary research field aiming to find patterns connecting affective states with its indicators, we can

¹Although some parts of this dissertation consider affect in its broader meaning.

define *emotion recognition* as a problem of finding links between input consisting of some observable indicators of a person’s internal state and target emotional states.

Some contents of this chapter originate from a larger article published in a peer-reviewed journal:

- [16] S. Saganowski, B. Perz, A. G. Polak, and P. Kaziemko, “Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review,” *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1876–1897, 2023. DOI: 10.1109/TAFFC.2022.3176135.

In the original article [16], descriptions of topics related to physiological signals (referenced in Sec. 2.2) were mainly inspired by Prof. Adam Polak with support from the dissertation’s author and his supervisors. All other sections in this chapter (and in the original article [16]) should be treated as our own work unless otherwise noted.

2.1 Emotions

Although emotions are not observable per se (one cannot enter another person’s mind), for years, researchers focused their efforts on trying to measure and understand them by creating various theoretical models, and observing when and how emotions influence our state and impact our daily lives.

2.1.1 Theories of emotion

Emotions are complex psychophysiological states involving interplays between internal feelings, external behaviors, and physiological reactions. They shape our perception, preferences, reactions, and relations, influencing our conduct and decisions, both minor and major ones [3, 4]. Because of their vital role, psychologists approach them very seriously and have yet to come to a conclusion about their nature. However, researchers trying to explain the character of emotions created some theories and perspectives regarding them.

One of the most well-known and influential theories is basic emotion theory. Models proposed by researchers like Ekman [39] or Izard [40] postulate the inherence of basic emotional states, biologically innate and thus present in all cultures. Most models assume six such basic states, namely anger, disgust, fear, happiness, sadness, and surprise, with slight variations possible. These emotions are believed to be unique and relatively constant in their causes, functions, and expressions. It implies

that emotions serve purely evolutionary purposes, such as reinforcing behaviors beneficial to social life (happiness) or triggering a fight-or-flight response (fear).

Appraisal theory provides a different view on the process of emotion induction. Proposed by Richard Lazarus [41], this theory posits that people first interpret events and situations regarding their goals, desires, and well-being, and emotions are a product of such appraisal. Contrary to basic emotion theory, it implies that external stimuli do not determine emotions in a distinct manner but are subject to interpretation. Therefore, the same event may evoke different feelings, depending on the situation or the person conducting the affective evaluation.

Next, we can outline the group of psychological construction models. They posit that emotions and other mental states are actively constructed and modified in the mind in reaction to sensory inputs instead of being predefined and hard-wired in the brain. These models emphasize the role of experiences and cognitive processes, such as memory and attention, in shaping emotion and indicate that variables like context, culture, knowledge, beliefs, perception, and other personal factors all influence the emotional response. It implies that generated mental representations of reality, and therefore affect, can change over time as a result of gaining new knowledge and experiences.

Finally, models that bring special attention to social and cultural factors are called social constructivist models. They are similar to psychological construction models in their assumption that emotions depend on an individual who interprets reality and gives it meaning. However, social constructivist models mainly focus on social norms, roles, and interactions in creating affect, while psychological construction models focus on cognitive processes and constructed mental representations.

2.1.2 Models of emotions

While theories from previous sections provide valuable insights into the processes that lead to emotion and the emotional experience itself, it's important to recognize their limitations. These theories, by themselves, are not sufficient to accurately measure and predict a person's emotions. Researchers have addressed this challenge by developing specific models within these theories, which could be used to study emotions by conducting experiments and analyzing obtained data.

One of the common ways of classifying emotional models is their division into two basic types: discrete and dimensional. Discrete models assume emotions that are precisely defined and independent from each other. A person may experience one

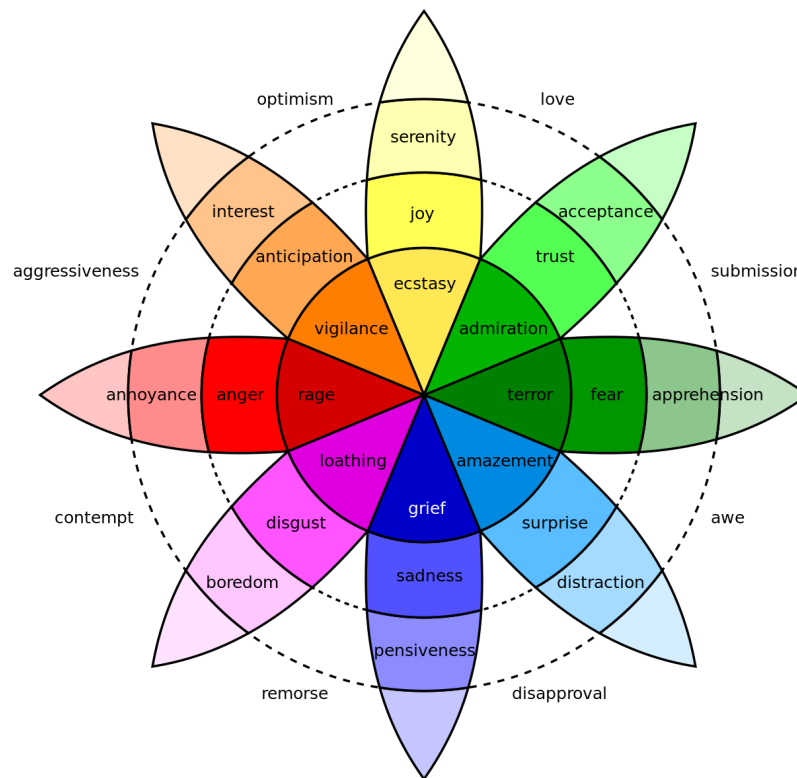


Figure 2.1: Plutchik's wheel of emotions (from [42]).

or many such states simultaneously, each with varying intensity. These assumptions restrict the number of possible emotions that can be felt, but models differ in the amount and naming of these states. Examples of well-known discrete emotion models include Ekman-Friesen [39] (*anger, contempt, disgust, fear, happiness, sadness, and surprise*), or Plutchik's wheel of emotions [45] (Fig. 2.1).

On the other hand, multi-dimensional models of emotion assume several orthogonal dimensions, each describing different component of affect using continuous values. The most well-known are the 2-dimensional Circumplex (*valence-arousal*, Fig. 2.2) [46] and the 3-dimensional PAD (*pleasure-arousal-dominance*, Fig. 2.3) [47] models. Typically, *arousal* denotes affective intensity or energy felt while experiencing emotion; *valence/pleasure* – emotional pleasure or polarization, a range from sad to happy; and *dominance* a sense of control over the situation, useful when differentiating between, e.g., states of anger and fear.

As all such models try to explain affect, we can move from one model to another, even between discrete and dimensional models. Dimensional models are often

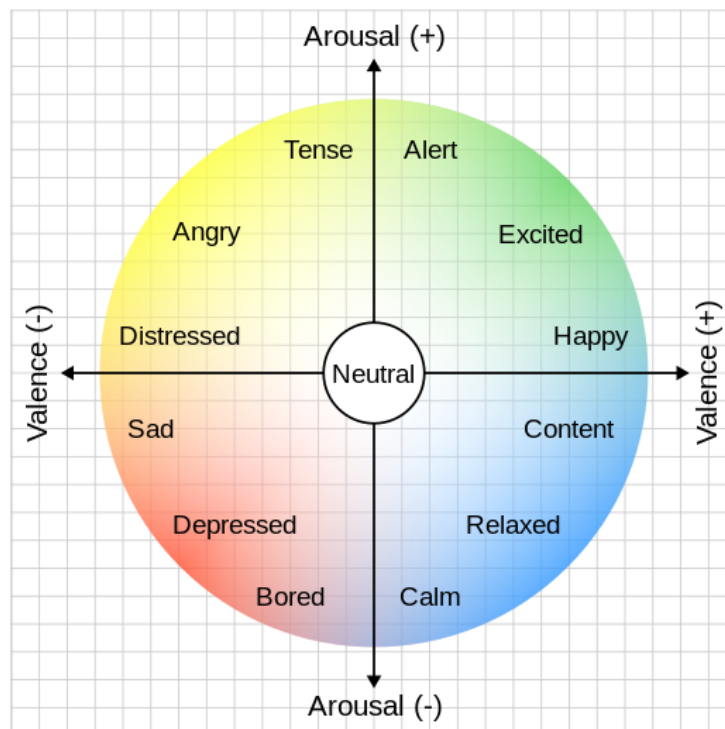


Figure 2.2: Circumplex model of emotions (from [43]).

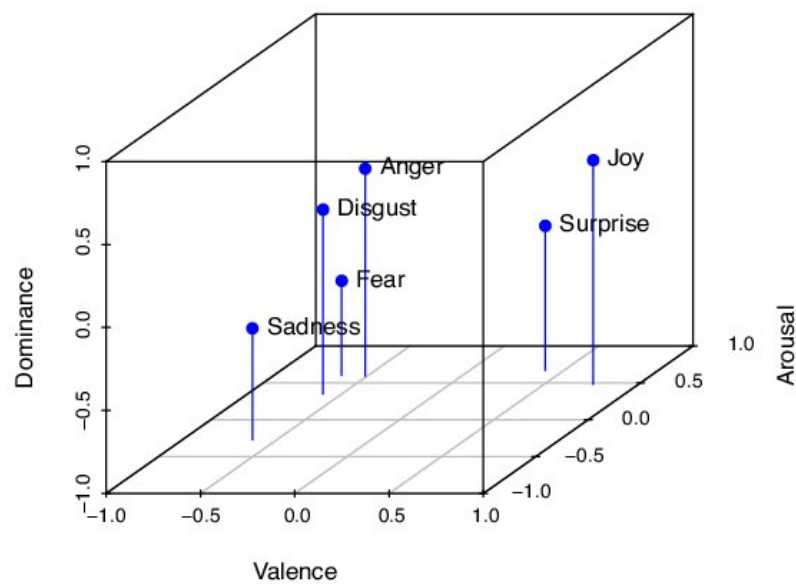


Figure 2.3: Pleasure-Arousal-Dominance (PAD) emotion model (from [44]).

described using discrete emotions to make understanding them easier, as terms such as *arousal*, *valence*, or *dominance* are rarely used in everyday language. Moreover, Russel proposed the Circumplex model [46] after asking people to place 28 discrete emotions in different positions relative to each other and analyzing their responses.

2.1.3 Communicating emotions

Suppose we want computers to employ emotion intelligence in a way that is natural to humans. In that case, we have to design them to use primarily bodily reactions to changes in emotional states (or *sentic modulation* as Picard [10] calls it), instead of relying on people explicitly naming their feelings. Not only would it be unnatural always to describe how we feel, but also, not many people can articulate their emotions well.

Some types of emotional expressions are easily visible to others and (often sub-consciously) serve as primary means of communicating emotions in everyday life. They are so important in everyday communication that some emotion theorists posit the existence of unique somatic response patterns, characteristic of emotion (e.g., unique facial expressions [48]). Other signals, such as changes in physiological signals, are hard for another person to perceive, but machines can easily perceive them using specialized sensors (e.g., heart rate from a smartwatch). Emotion recognition systems could utilize signals both easily and hardly (if at all) apparent to other people (Tab 2.2), or different combinations thereof [16, 49, 50].

Table 2.2: Categories of emotional cues for affective research (based on [10]).

Observability by other people	Signals revealing emotional state
Apparent to others	<ul style="list-style-type: none"> • Facial expressions [49, 51] • Voice and speech (e.g., intonation) [52] • Body language and gestures (e.g., movements, posture) [49] • Gaze and eye movements (e.g., pupillary dilation, gaze aversion) [49, 53] • Written text [54]
Hardly or not apparent to others	<ul style="list-style-type: none"> • Physiological signals (e.g., respiration, heart rate, perspiration, blood pressure) [16, 49]

2.1.4 Applications of emotion recognition

Near the end of the 20th century, people started to realize the potential of affective computing. One of the first use-cases mentioned in the literature, and still valid to this day, is to use affect recognition for improving human-computer interaction [10, 55–57], e.g., by adapting interfaces or system's reactions to better suit user's emotions, e.g., by adjusting game difficulty or design [58, 59], or providing better recommendation of music or movies [60]. Over the years, researchers have been finding new possibilities for utilizing systems capable of recognizing affective states. Such systems could help people suffering from panic attacks [61], increase people's resilience and productivity [62], assist teachers in the classroom [63], help doctors diagnose patients and provide custom treatments [64], improve people's wellbeing and longevity [65], or even support the emotional development of children with autism spectrum disorder (ASD) [66, 67].

However, computerized emotional intelligence could be used for more than directly improving people's daily lives. Contrary to what many people believe (or would like to believe), emotions play an essential role in decision-making (also the rational one), perception, and other cognitive functions [2–4]. Emotional machines could benefit from human-like decision-making processes. Learned emotional biases (or so-called "gut feeling") could be used by computers in situations where available options cannot be fully explored, either due to the nature of the problem or high computational cost [10, 68]. It could lead to good results in certain situations, although not always would they be explainable.

Seeing such a vast range of benefits, one could easily forget about the possible dangers that affective computing introduces. Data about people's emotions could be used against them, especially when emotion recognition systems become more accurate [69]. Inaccurate predictions about emotions, on the other hand, could lead to misunderstandings and possible harm for a subject of recognition, such as career opportunity loss in a work environment or mental health degradation in a social situation [70]. For example, incorporating facial emotion recognition (FER) algorithms, known for their questionable quality and reliability [18, 71], into systems monitoring people at work, could lead to severe consequences based solely on wrongful predictions. Also, in a more sinister scenario, some totalitarian state could utilize universally adapted emotion recognition to punish troublesome individuals or to control the populace and its feelings [10].

2.2 Physiology

Biological indicators of emotions [8] have been a topic of research for years. In 1954, Schlosberg asserted that electrical skin conductance is a reliable measure of the intensity of emotional arousal [72]. Ekman, Levenson, and Friesen demonstrated during the 1980s and 1990s that the autonomic nervous system's response to deliberately produced emotions can be discerned through physiological signals [7, 73]. Specifically, they identified correlations between six primary emotions and metrics, such as heart rate, finger temperature, and skin conductance, attributing these connections to the functional specificity of the autonomic nervous system (ANS). These associations serve as the primary driver for the affect recognition from physiological signals.

2.2.1 Biosignals

Physiological systems and biosignals

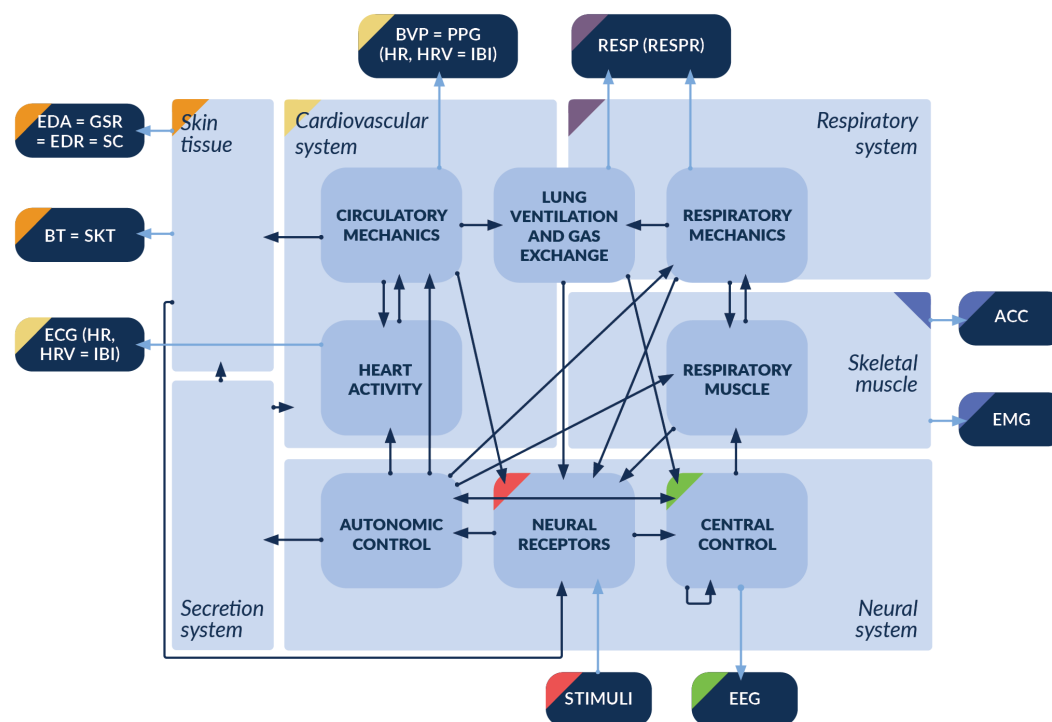


Figure 2.4: Interrelationships between physiological systems and biosignals. From Saganowski et al. [16].

The human body can be seen as a sophisticated system, dynamic and nonlinear in its nature, characterized by complex feedback mechanisms between individual organs and entire physiological systems. It is designed to maintain internal stability (homeostasis) and orchestrate appropriate responses to changes in the external environment (see Fig. 2.4). In the presence of external stimuli, signals passed from

sensory organs to the brain cause the activation of the central nervous system (CNS). The processing of information about stimuli is mainly done by the cerebral cortex. This activity can be monitored using the electroencephalogram (EEG) and seen as distinct brainwave patterns across EEG frequency subbands (delta, theta, alpha, beta, and gamma).

The CNS, together with the peripheral nervous system (PNS), predominantly regulates skeletal muscles, including the respiratory ones. Its activity can be monitored using muscle electrical activity via electromyogram (EMG) or body movements via accelerometers (ACC). The same muscle activity can interfere with measurements of other biological signals, especially electrical ones, e.g., EEG or electrocardiogram (ECG). The action of the aforementioned respiratory muscle impacts airflow, thus shaping the respiratory waveform (Resp) and governing gas exchange in the lungs. It determines oxygen saturation and blood pH, which are monitored by chemoreceptors and analyzed by the autonomic nervous system (ANS) and the central nervous system. Nervous systems respond by regulating cardiac activity, causing depolarization and repolarization of atria and ventricles, prominently represented in the electrocardiogram (ECG) waveform and its QRS complex showing the ventricles' depolarization. Heart rate and stroke volume dictate the pulsatile blood flow. Together with arterial smooth muscle tone, controlled by the ANS, they also influence temporary blood pressure. These effects can be monitored noninvasively using photoplethysmogram (PPG) or blood volume pulse (BVP) signal.

The ANS is also responsible for modulating airway smooth muscle tone and impacting respiratory mechanics. This relationship can be seen, for example, in a sudden lack of breath during a panic attack. Additionally, blood flow intensity influences the regulation of body temperature and hormone transport speed, reflected in measures such as body temperature (BT), skin temperature (SKT), and electrodermal activity (EDA), characterized by tonic skin conductance level (SCL, slower) and phasic skin conductance response (SCR, faster).

2.2.2 Signal processing

When working with physiology, signal preprocessing is an essential first step of any analysis. It is essential to remove noise that contains no information on emotional reactions but is present in data, thus obfuscating affect-relevant information. Such distortions originate from both external and internal sources, such as (1) external electromagnetic fields influencing measurements, (2) noise generated inside elec-

tronic circuits, (3) movements of the body and changes in positions of sensors on the body surface, (4) overlapping of signals from different organs (e.g., brainwaves and facial muscles), (5) or temporary malfunctions of sensors. The methods used to retrieve the original form of the signal are either based on its known properties or are data-driven.

Before converting the voltage a sensor registers into a digital signal, it is necessary to perform initial anti-aliasing filtering (Nyquist-Shannon sampling theorem). Next, further filtration is typically used (analog or digital) with additional smoothing (reducing high-frequency components) to emphasize desired frequency components and limit others. To distinguish desired and undesired components, one may use decompositions, e.g., wavelet transform representing a nonstationary signal using different scales related to frequencies. Another tool is independent component analysis (ICA), which can be used to distinguish independent source signals that overlap while being recorded by different sensors.

Normalization is often used to ensure comparable levels of signal energies and extracted features, especially when data is collected using different devices. If extreme values are present in the data, one can consider using winsorization to reduce their effect. Finally, if the signal consists of disrupted fragments, interpolation may be used to replace altered fragments with values estimated using statistical properties of adjacent data.

2.2.3 Signal features

Many studies follow the classical approach to machine learning, requiring extraction and selection of hand-crafted signal features. Such features represent the specific properties of analyzed data and are often extracted using a sliding window over the signal. Such features are primarily computed within three domains: (1) time, (2) frequency, and (3) time-frequency or time-scale (for nonstationary signals). The most commonly used transformations include Fourier and wavelet transforms, along with the decomposition of EDA into tonic and phasic components.

Additionally, specific scalar metrics are computed from these signals or transformations, resulting in the ultimate set of extracted features. These metrics encompass various categories, including (1) morphological properties, (2) dynamic properties as defined by Hjorth parameters, (3) energetic parameters such as root mean square (RMS) and power spectral density (PSD), as well as (4) statistical indices like mean value, median, and standard deviation (SD). Moreover, due to the nature of phys-

iological signals, measures for nonlinear systems are often employed to represent them, such as Poincare plots or entropy.

After computing descriptive features, some of them may be irrelevant. Often, only their subset is selected for experiments to reduce the dimensionality of the feature space. If done correctly, feature selection helps increase the efficiency and performance of a machine learning model. As testing all possible combinations of features is usually infeasible, researchers use a few schemes for feature selection. The first is transformation, where features are projected onto some arbitrary space, using, e.g., PCA, and later selected. Another one is filtering, which uses a criterion such as information gain (IG) and a threshold to select relevant features. Wrapper methods use a classification algorithm (e.g., SVM) as a proxy to check features relevancy for a given task (for instance, classification of emotions). Embedded methods rely on the ability of deep learning algorithms to process the input in a nonlinear fashion and generate alternative feature spaces. Typically, the above approaches are complemented by adding the most beneficial features (sequential forward selection, SFS) or eliminating the least favorable ones (sequential backward selection, SBS).

Chapter 3

CRITICAL LITERATURE REVIEW

The previous chapter introduced knowledge essential for understanding this dissertation and the broader field of affective computing. In this chapter we present the results of our systematic literature review that was focused on methods suitable for real-life emotion recognition, i.e., experiments that were either conducted in the wild, or were done in the controlled setup that resembles real life.

When designing emotion recognition systems for real life, one needs to consider many factors impacting its capabilities. Some of them are well-known to researchers, as they also appear in laboratory studies. Others are unique to uncontrolled environment and, therefore, new for affective computing scientists. With this in mind, we not only describe methods used in affective computing for real life. We also highlight similarities and differences between experiments conducted inside and outside the laboratory by contrasting them across several study components that we identified.

Contents of this chapter originate from the co-authored article, published in a peer-reviewed journal:

- [16] S. Saganowski, B. Perz, A. G. Polak, and P. Kazienko, “Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review,” *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1876–1897, 2023. DOI: 10.1109/TAFFC.2022.3176135.

In the original article [16], descriptions of topics related to data processing were mainly inspired by Prof. Adam Polak with my and my supervisors’ support. All other sections should be treated as our own work unless otherwise noted, with the special contribution of myself to descriptions of the used methodology, participant-specific data, emotional stimuli, context affecting emotion experience, emotion self-assessment and labeling, windowing and learning case definitions, and utilized reasoning models.

3.1 Methodology

This literature review belongs to the category of *Systematic Literature Review* [74] (SLR), and its main task was to answer the following question:

Can wearables be used to recognize emotions in everyday life?

We researched the question using three databases of scientific literature, namely Scopus, Web of Science, and Google Scholar (via Publish or Perish). The search terms were designed to capture a wide range of literature related to the intersection of emotions and wearable technology. We asked the following query:

[emotion or affective] and [wearable* or (smart watch) or iot or (personal device*) or (ambient intelligence) or (smart device*) or (smart band*)].*

The search was performed in three iterations on August 30, 2019, August 30, 2020, and March 4, 2021. The first round resulted in 2,993 records found, out of which 2,384 remained after removing duplicates, patents, and non-English resources. In the second search, we limited the investigated time span to years 2019 and 2020 (to skip articles that we already knew of), and obtained 577 papers, out of which 313 were left after initial filtering. During the third round, we restricted results to the years 2020 and 2021. It returned 553 papers, out of which 327 qualified for further review. Additionally, 27 papers were added manually after reviewing references in relevant articles and analyzing other works authored by the same researchers as the already included studies. In total, we evaluated 3,051 papers using the following set of inclusion and exclusion criteria to compile a set of articles relevant to our question:

Inclusion 1 Personal devices/wearables were used to recognize (classify) various emotions. Device/wearable should enable emotion recognition in everyday life;

Inclusion 2 Personal device/wearable was described, or the description was available elsewhere;

Inclusion 3 At least one physiological signal was monitored and utilized to emotion recognition;

Exclusion 1 The study was performed on a population less than five subjects;

Exclusion 2 Only a single emotion or its levels was considered;

Table 3.1: A number of articles excluded from SLR with respect to particular inclusion (In) or exclusion (Ex) criteria (from [16]).

Criterion	In1	In2	In3	Ex1	Ex2	Ex3	Ex4
No. of excluded articles	857	20	265	99	181	113	118

Exclusion 3 None of the exploited devices was personal/wearable/ portable;

Exclusion 4 The device had modules interconnected with cables, e.g., BioPac system where sensors were wired to the development board.

Out of all reviewed articles, 1,398 studies were discarded solely based on their titles and abstracts, as their irrelevancy was evident. A summary of the other excluded articles rejected due to particular SLR criteria can be found in Tab. 3.1. For inclusion criteria, we report the number of papers failing to satisfy the particular statement. Numbers for exclusion denote how many papers were removed because of a given criterion.

When excluding articles, we focused on solutions, systems, and devices that are applicable across all daily-life situations, rather than specific scenarios. No constraints should limit their applicability, e.g., webcam-based emotion recognition systems, as they require subjects to be seated in front of the camera, or articles utilizing only EEG signals, as existing brain-measuring devices are too susceptible to everyday-life noise to be applicable outside the laboratory. Even small movements can alter recorded brainwaves due to (1) electrical signals caused by muscle activity, (2) movement of electrodes, or (3) activation of different brain areas overlapping each other. Moreover, we focused only on emotions, i.e., affective states spanning and influencing physiology only for a short time [8], as literature considers long-lasting states as moods rather than emotions [34].

We also excluded articles focusing only on a single emotion, e.g., only *anger*, as they do not provide insights about recognizing different emotions, or only arousal as by itself arousal does not correspond to any emotion. On the other hand, we included articles recognizing only valence levels, as this dimension divides the emotion spectrum into positive and negative states, e.g., happy vs. sad. At last, we discarded articles that did not perform emotion recognition, e.g., focusing solely on correlations obtained from statistical analysis [75]. As they did not provide any reasoning about the viability of emotion recognition, they were also irrelevant to our

research question. A total of 34 articles passed the above process and were included in the SLR.

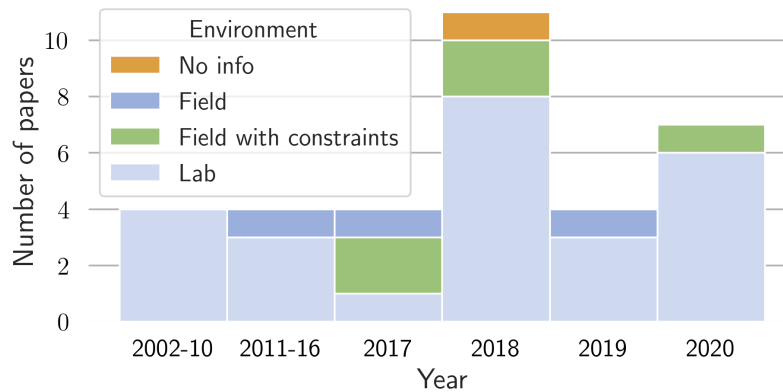


Figure 3.1: Number of papers in relation to the year of publication and study environment (from [16]).

We noticed that the number of publications passing our criteria was growing during the last five years of the search, Fig. 3.1, with the oldest relevant article published in 2002 and eight papers published by 2016. Also, a significant increase in relevant articles is visible starting in the year 2017, with 11 articles published solely in the most fertile year – 2018. Moreover, the research environment changed across the years, as no field study was conducted between 2002 and 2015, and at least one field study a year since 2016, when Exler et al. [76] carried out their research outside the laboratory. Out of those studies, five took place in a constrained environment ¹. A total of 25 studies were conducted in the laboratory, eight studies took place in real life, and in one paper, the authors failed to specify the environment. In our opinion, this shift is driven by the need for solutions suitable for real-life setups and enabled by improvements in wearable technology and sensors over the years, making measuring devices more practical, portable, and convenient. Unfortunately, no article published between January 1 and March 4, 2021, passed our research criteria.

3.2 Emotion recognition study design

When studying the designs of studies focused on emotion recognition, it is hard not to notice some commonalities between them. We identified nine such research stages,

¹We derived from the description that Dao et al. [63] measured emotions only in a classroom, i.e., a field study with constraints.

Research stages

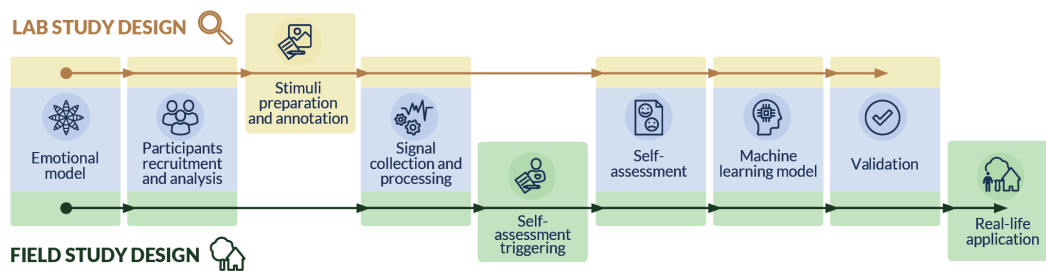


Figure 3.2: Common and unique research stages (study design) for emotion recognition in the lab and in the field (from [16]).

similar between examined studies (Fig. 3.2). However, each of the components may require different implementation depending on the target environment, as laboratory and field studies differ in terms of (1) experienced emotions, (2) emotional stimuli, (3) labeling procedures, (4) acquisition of emotion assessments and (5) physiology, and (6) other factors such as the amount of collected data or battery life problems (Tab. 3.2). Therefore, designing an affective study requires careful consideration of each research stage.

First is a decision about the emotional model assumed. It influences other components, in particular the choice of stimuli and self-assessments, and shapes the final reasoning and conclusions. Also, the wrong choice of the emotional model may result in an applicable but complicated procedure, e.g., finding stimuli eliciting two emotions - happy vs. sad is easier than finding stimuli for 27 different emotions, which are also likely to co-occur.

Second is the recruitment and training of study participants. Since people may have conditions that influence collected measures, like heart problems disrupting cardiac signals, researchers have to prepare appropriate exclusion criteria. For subjects that pass preselection for a lab study, a simple explanation of the experiment is usually sufficient, as researchers set up the experiment and can assist subjects during the procedure. Conversely, in field studies, participants have to be trained to put on measuring devices, manage them, and solve simple problems arising during the study. Additionally, researchers should always be available for help (e.g., via phone call, ideally 24/7) [16].

The next component concerns collecting physiological data with wearables. In the laboratory, researchers typically utilize precise and reliable medical-grade equipment. In addition to devices that provide high-quality data by themselves, artifacts

Table 3.2: Main differences between laboratory and real-life emotion recognition studies. Advantages are marked with '+', disadvantages with '-', and points falling in-between with '±' (from [16]).

Category	Laboratory study	Real-life study
Emotions experienced	<ul style="list-style-type: none"> – In a controlled environment – Impacted by unnatural conditions – Limited by chosen stimuli 	<ul style="list-style-type: none"> + In natural environment + Full range of emotions – Difficult to capture
Stimuli	<ul style="list-style-type: none"> ± Well-known, fully planned and controlled + No distractions or interference with other situations + Known context 	<ul style="list-style-type: none"> + Real, immersive – Unknown / uncontrolled – Reaction possibly impacted by context or life conditions (drugs, fatigue)
Labeling (ground truth)	<ul style="list-style-type: none"> + Self-assessment + Expert-annotated stimuli + Observed and derived by external experts 	<ul style="list-style-type: none"> – Mainly self-assessment + Observed by a nearby person (relative, friend)
Self-assessments	<ul style="list-style-type: none"> + Detailed + In precisely controlled moments 	<ul style="list-style-type: none"> – Limited in scope – Sparse – Response usually delayed [76]
Measuring physiology / devices	<ul style="list-style-type: none"> + Precise, medical-grade devices, giving high-quality data + Unrestricted number and type of devices + Small number of artifacts (stationary position, well-known conditions) 	<ul style="list-style-type: none"> ± Personal, convenient wearables, giving signals of lower quality [77] – Only few devices feasible + Convenient and unnoticeable measuring – Artifacts from movements and in-the-field conditions
Additional factors	<ul style="list-style-type: none"> + Static environment (temperature, lighting, etc.) + Participants require only basic training + No problems with battery life or data synchronization ± Results in relatively small amount of data 	<ul style="list-style-type: none"> – Device charging and data synchronization required – Participants require extensive training – Hotline to technical support required ± Results in a large amount of data

are further limited by the stationary position of a subject during an experiment. These methods are not suitable for in-the-wild studies where people are performing all kinds of activities, and it is best when devices are unobtrusive and comfortable to wear. Unfortunately, sensors in such devices are of lower quality [77], often resulting in inaccurate and noisy signals. Additionally, these studies require more engagement from subjects because of additional responsibilities, e.g., charging devices, putting them on according to some instructions, or uploading their data to the cloud.

Concurrently with physiological signals, their emotional annotations have to be collected. Triggering questionnaires in the laboratory can be done based on consumed stimuli, usually right after they end. Moreover, questionnaires can be designed to match selected stimuli and expected reactions precisely. In real life, in turn, triggering self-assessments poses a great challenge. Accurately determining moments of emotional reactions is challenging, and researchers typically utilize self-triggered or at-random questionnaires². Additionally, because emotional moments are the most valuable in such research, utilized questionnaires should be short and simple to make filling them out easy for participants experiencing strong emotions [76].

Collected signals have to be processed to allow the creation of reasoning models. These signals or features describing them are next combined with ground truth labels and used to model the psychophysiology of emotions. To create reliable reasoning models and make them ready for use in real life, researchers should carefully consider employed validation methods, and ideally perform hyper-parameter optimization. Reasoning models created using data from laboratory experiments may achieve good accuracy on data collected in another controlled study but prove useless when applied outside the lab. On the other hand, models created on lower-quality data from the field may exhibit lower accuracy on a single task, but better generalize between different real-life conditions.

Reviewing included articles also revealed seven major scenarios for emotion recognition (Fig. 3.3). The first five are the in-the-lab scenarios, whereas the last two refer to outside-the-lab research. The main differences between them are (1) labeling strategy, i.e., obtaining emotional label (ground truth), and (2) stimuli, i.e., emotion elicitors. Identified scenarios were used as follows ('?' denotes that it was not provided by authors but deduced by us):

²Later in the thesis (Chap. 7) we describe experiments conducted with pre-trained models for emotion detection that may help in addressing this issue.

Emotion recognition scenarios

Legend: Scenario-specific component

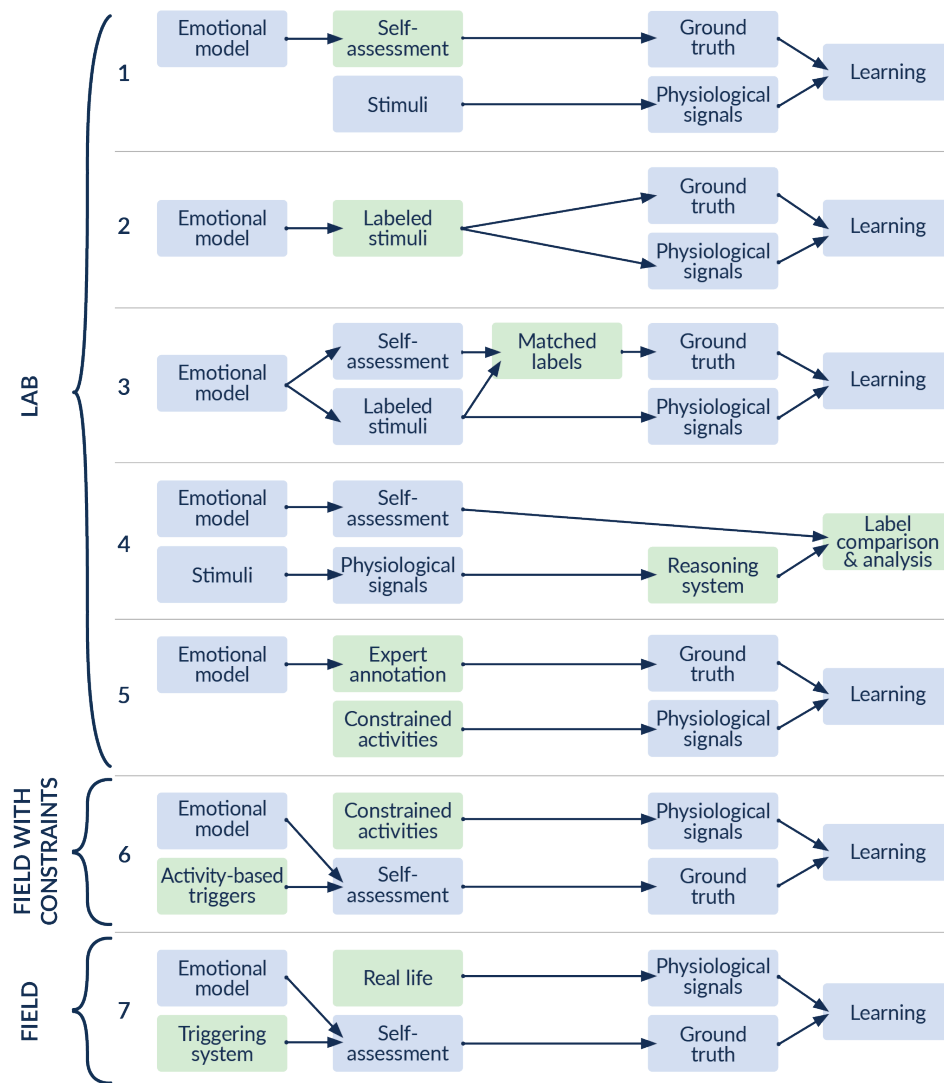


Figure 3.3: Emotion recognition scenarios identified in SLR (from [16]).

Table 3.3: Participants' metadata surveyed in studies (from [16]).

Characteristics	References
Age	[11, 76, 78–86, 88–94, 96–104, 106, 107]
Gender	[11, 76, 78–85, 87–94, 96, 99–101, 103–107]
Profile	[11, 63, 76, 78, 79, 82–91, 93, 99, 100, 102, 105, 107]
Health condition	[11, 78–80, 82–84, 92, 93, 97–101]

- **Scenario 1** was used in 11 articles [78–86], and probably in [87]?, [88]?
- **Scenario 2** in 7 articles [80, 89–93], [94]?
- **Scenario 3** in 7 articles [95–100], [87]?
- **Scenario 4** in 1 article [101]
- **Scenario 5** in 2 articles [102], [88]?
- **Scenario 6** in 4 articles [63, 103–105]
- **Scenario 7** in 4 articles [11, 76, 106], [107]?

Unfortunately, we were unable to identify a scenario in one article [108]. Further details on individual papers, such as used emotional models or machine learning problems defined from these models, can be found in Tab. A.1.

3.3 Study participants

At first thought, participant recruitment does not seem like a hard task – advertise the study, and some volunteers will come. However, even similar individuals (e.g., all of one nationality) may differ in their emotional responses to the same situation or stimulus, e.g., women are often perceived as more emotional than men [109], with their reactions being more intense [110]. Moreover, it was emphasized that gender differences in any particular modality of emotional expression are culturally and situationally specific [111]. Such response-impacting factors are not limited to gender, and they include biological or cultural traits such as age [112], subject's profile [113] and their health [79, 82] (Tab. 3.3).

Among many factors that can influence study results, age was the one reported most often (Tab. 3.3). In the studied articles, most of them focused on a relatively young population, two on a group of people who were 28 years old on average [103, 104], and 20 on people between 18 and 40 years of age. In two studies, subjects could be older than 40 [82, 101], and one study focused solely on infants and toddlers (up to 30 months old) [102].

Another usually reported factor is the gender of participants. In 12 papers, the subjects were mostly or exclusively males; in 12, they were mostly or exclusively females, and in three genders of participants were balanced. In 20 articles, participants were described in more detail, i.e., their occupation, nationality, or other group that they identified with. Often (14 papers), the experiments were conducted at least partially on students. Subjects' ethnicity or nationality was reported in four works [82, 89, 90, 93].

Subjects' condition was studied in 14 papers, with researchers requiring participants to be in good health [11, 78, 80, 92, 97, 98, 101]. Conditions that were examined in more detail include mental health or emotion understanding problems [11, 82, 84, 92, 93, 97, 98, 101], cardiovascular diseases [79, 84, 101], neurological diseases [79, 83], vision difficulties or its correction [82, 99, 100], or excessive sweating [101].

3.4 Temporal context affecting emotion

Features such as gender, age, ethnicity, or disease are not the only ones that can impact the results of a study on emotion. Other factors shaping emotional experiences and physiological responses, like specific drugs or activities, often have more immediate effects. Therefore, we name them temporal context.

Although the temporal context is known to be important, only ten studies collected data about it (Tab. 3.4). Participants were asked about the use of medication or non-medical drugs (e.g., alcohol, caffeine, tobacco) in six studies, and in all of them, subjects had to be free from these substances. Similarly, Zhao et al. [79] required that experiment subjects refrain from consuming caffeine, fat or salty food for an hour before the experiment; Shu et al. [92] asked people to abstain from alcohol and medications for 72 hours before the study; and Ragot et al. [82] subjects did not use medication on the experiment day. Dao et al. [63] and Schmidt et al. [11], who performed outside-the-laboratory studies, collected data about subjects' lifestyle and activities [63], the intensity of physical activity, and sleep quality [11].

Specific categories of temporal context should be considered in relation to the

procedure employed in the conducted study. Substances or activities affecting subjects' perception are especially important if researchers focus on self-reported subjective emotions, usually collected using self-assessment questionnaires. On the other hand, if emotions are labeled extrinsically, annotations are not strongly impacted by the individual context. However, even if psychological assessments are not affected by the context, people's physiology may still be affected, making the acquired data useless. Also, the two types of context are often affected together, e.g., caffeine may increase a person's arousal [114], heart rate, and blood pressure [115].

In the laboratory, it is easy to detect and control temporal context. Not only may researchers ask about the amount of time since the last ingestion, but it is also impossible for subjects to consume any substances without the researcher's knowledge. This way, when analyzing signals from in-the-lab experiments, it is possible to measure and account for the context. Field studies are fundamentally different in this aspect, as it is practically impossible to force people to refrain from consuming various substances, and gathering contextual data is much more difficult due to subjects forgetting to report relevant information or having trouble providing exact data on the context, e.g., time that passed since drug consumption.

Out of the reviewed studies conducted in real life, only two papers collected contextual information. Schmidt et al. [11] gathered physical activities and sleep quality data, but they did not account for context at the reasoning stage. Dao et al. [63] did not clearly state that they utilized data on physical activities for recognizing emotions, but they stated that they used it for finding lifestyle-mood patterns that were sometimes associated with emotions. Hu et al. [88] and Majumder et al. [94] approached context differently and decided to create it by making participants sit or walk during the experiment. Although neither of them compared these approaches, performing such studies and searching for differences in emotions between various contexts may be a good idea.

Almost every reviewed paper (91%) included some information about people's traits that may influence emotions (e.g., age, health condition); only about a quarter of studies considered temporary context (e.g., drugs taken), and none accounted for them during analysis. If utilized at all, questionnaires about different factors were used only to exclude participants who took undesired substances.

Table 3.4: Context surveyed or considered in reviewed studies. Only [11, 63] are field studies (from [16]).

Context	Details: used by
Medication	[78, 79, 82, 92, 101]
Food	[79]
Drugs	alcohol: [84], caffeine: [79], tobacco: [84]
Lifestyle	sleep quality: [11], activity: [11, 63]
Activity	rest vs walking: [88, 94]

3.5 Collecting physiological signals

In this dissertation, we focus on research where collected measures involve both physiological signals and questionnaires on emotions. For monitoring physiology, most in-laboratory setups utilize medical-grade devices, giving very high-quality signals. To be suitable for real life, devices measuring physiology have to be in the form of wearables - they have to be comfortable to wear and inconspicuous. However, signals recorded with such devices are of lower quality than their in-the-laboratory counterparts [77] because of various artifacts and other daily mistakes, such as forgetting about charging devices or synchronizing data (Tab. 3.2).

The most popular physiological signals that can be collected in real life and used for emotion recognition are electrodermal activity (EDA), body temperature (BT), photoplethysmography (PPG), and body acceleration, (Tab. 3.5). For lab studies, BioPac MP160 or ProComp Infiniti can be used, but such devices are usually large and sophisticated and use a lot of wired connections, making them unsuitable for real life. However, there are more than 50 wearables that measure signals useful in emotion recognition and are small enough to use them in real-life studies [116]. In this review, we focused on studies using such real-life ready tools (Tab. 3.6).

Among the devices chosen in the reviewed papers, Empatica E4 was the most popular due to its many advantages, such as high-frequency sensors, raw signal availability, long battery life (compared to others), and simple-to-use API. In stationary conditions, signals provided by Empatica have quality comparable to the ones provided by ambulatory devices, but they are also prone to motion artifacts [117]. It also has several downsides, the major ones being its high cost, lack of non-measuring features, and relatively non-attractive appearance. Additionally, Borrego et al. [118] investigated the EDA signal collected with Empatica E4 and obtained much worse

Table 3.5: Physiological signals used for emotion recognition (from [16]).

Physiological measure	References
Electroencephalography (EEG)	[81, 85]
Electrocardiography (ECG)	[76, 88, 95–98]
Electromyography (EMG)	[11, 83]
Electrodermal activity (EDA) or response (EDR) / Galvanic Skin Response (GSR) / Skin conductance (SC)	[11, 63, 78–80, 82–87, 89, 90, 93, 94, 101–105, 108]
Photoplethysmography (PPG) / Blood volume pulse (BVP)	[11, 63, 79, 81–83, 86]
Body temperature (BT) / Skin temperature (SKT)	[11, 63, 78–80, 83, 84, 87, 89, 90, 94, 101, 103, 104]
Respiration rate (Resp)	[11]
Acceleration (ACC)	[11, 63, 99, 100, 103–105]

Table 3.6: The most popular physiology-measuring devices in SLR (from [16]).

Device	Type	Sensors and signals	Used by
Emotiv Insight	Headband	EEG, ACC, GYRO, MAG	[81, 85]
Empatica E4	Wristband	PPG/BVP, EDA, ACC, SKT, tags	[11, 63, 78–82, 84, 86, 91, 94, 101]
Microsoft Band 2	Smartband	PPG / BVP, EDA, ACC, GYRO, SKT, BAR, ALT, AL, UV, STP, CAL, UV	[83, 85, 93, 103–105, 108]
BodyMedia SenseWear	Armband	EDA, ACC, SKT	[87, 89, 90]

measures than with a laboratory-grade device.

Microsoft Band 2 smart band was another popular device, offering sensors similar to Empatica’s together with other functions, such as activity and sleep monitoring, smartphone integration, and watch functions. Also, it was found to give precise measurements in stationary conditions [119]. Simple EEG headbands, e.g., Emotiv Insight, could be very useful for precise emotion recognition, but because they are prone to artifacts and not very comfortable, their applications are limited to specific,

stationary setups – in the lab or at home.

Other devices included: Polar H7 [99, 100] or H10 [84], or ekgMove [76] chest straps; Samsung Gear 2 [99, 100] or Algorand F8 [92] smartwatches; smart clothing (XYZlife Bio-Clothing) [97, 98] or a self made device [88]; Q-sensors [102]; wristbands Silmee W20 [107] or Mio Link [106]; Wacom Bamboo Ink stylus, Shimmer GSR+, and PPG ring [84]; BodyMedia SenseWear [87, 89, 90]; Biopac BioNomadix MP150 [82]; or RF-ECG biosensor kit [95, 96].

With the improvements in sensor quality, the popularity of smartwatches should increase in the coming years, as they are the most ubiquitous and unobtrusive out of the above devices, can offer features other than signal measurement, and can collect multiple measures (acceleration of the human body, plethysmography, and sometimes electrocardiography or electrodermal activity).

3.6 Emotional models and adjustments

One of the most important decisions during the study design phase is the choice of the emotional model, as it impacts other parts of the study, particularly the selection of stimuli, self-assessment, and final reasoning. While current models of emotions can be applied to both lab and field studies, they may not be equally convenient. In lab studies, the outcomes of the experiment greatly depend on the chosen stimuli, and this choice is affected by the selected model of emotions. For example, choosing stimuli for eliciting two very distinct emotions, such as joy and grief, is simpler than invoking many different emotions, where multiple of them may co-occur during a single stimulation. On the other hand, a limited choice of stimuli results in an equally (or even more) limited number of emotions occurring during the experiment.

In real-life studies, daily situations act as stimuli. They are often unpredictable and invoke emotions that are more diverse and intense than in a controlled environment, but they differ between subjects and can be neither controlled nor studied in detail by researchers. Although a chosen model of emotions does not restrict experienced emotions, it dictates the questionnaires used in the study, therefore impacting the amount of collected information (registered emotions) and affective states that can be studied using the data.

Our analysis shows that most of the considered papers utilized discrete emotion models, often with custom emotion categories (Tab. 3.7). Researchers usually create a model of interest by modifying already-known models, e.g., by choosing a subset of five emotions from Plutchik's model [45]: *anger, fear, disgust, sadness,*

Table 3.7: Trigger times, types of self-assessment, and emotional models utilized in studies (from [16]).

Study Environment	Trigger type	Questionnaire type	Model of emotions	References
Laboratory	After stimulus	Standardized	Discrete	[78, 80, 95, 96, 99–101]
		Standardized	Dimensional	[82–84]
		Own	Discrete	[86, 87, 89, 90, 97, 98]
		Own	Dimensional	[79, 81]
	Time-dependent	Own	Dimensional	[85]
Constrained field	After stimulus*	Standardized	Dimensional	[105]
	Quasi continuous	Standardized	Dimensional	[103, 104]
	Voluntary	Own	Discrete	[63]
	Not specified	Own?	Discrete	[107]
Real life	At random, EMA, Voluntary	Standardized	Dimensional	[11]
	On events, Time-dependent, Voluntary	Standardized	Dimensional	[76]
	Not specified	Own?	Discrete	[106]

joy + neutral [95, 96], or by ignoring *surprise* state [106] from Ekman-Friesen model [39].

Most papers utilized models of four emotions such as *anger, fear, sadness, happiness* [97, 98, 120], with *neutral, relax, or peace* added [97, 98] as a fifth state. Other modifications included replacing one of these four emotions depending on authors' focus, e.g., replacing *anger* with *love* [101]; *fear* with *calm* [93], *pain* [78] or *relax* [107], and *happiness* with *frustration* [87]. Different four-emotion models consisted of *cheer, sadness, erotic, horror + neutral* [80] and *joy, sadness, stress, calmness* [108].

More emotional models included: (1) grouping emotions into classes – *positive* and *negative* [98] or *positive, negative, and neutral* [86]; or (2) choosing three or four

emotions from Circumplex model [46] with one of them treated as a neutral state: *joy, boredom, and acceptance* treated as *neutral* [102]; *happiness, sadness, and neutral* [99, 100]; or *anger, happiness, sadness, and neutral* [94]. Dao et al. [63] tested six discrete emotions, namely: *boredom, excitement, happiness, relax, stress, serenity*.

Saxena et al. [86] also considered reducing their initial four dominant emotions *anger, disgust, fear, and sad* into *negative* state and confronted it against *positive (amusement), anxiety and neutral* state.

Contrary to discrete models, dimensional emotional models assume the existence of multiple orthogonal dimensions describing affect (Sec. 2.1), with the circumplex (*arousal-valence*) model being the most popular. It occurs that the more complex nature of dimensional emotional models makes them less prone to modifications.

Out of articles utilizing dimensional models, in nine papers, the circumplex model was utilized [11, 79–84, 91, 108], and one paper used PAD (*pleasure, arousal, dominance*) model [105] without any modifications. In case of adjustments, the most popular approach was to utilize just the *valence* dimension [76, 92, 103, 104]. The only article introducing uncommon dimensions was written by Martens et al. [85] with *interest, energy, valence, focus, tension* used.

An interesting observation is that the same emotional model, e.g., Russell's Circumplex [46], can be used in its 2-dimensional [91] or treated as a discrete model with three [102] or four [107] emotions + neutral. Overall, the initial psychological models are often modified to suit the needs of researchers.

3.7 Emotion labeling

Before modeling human emotions, researchers have to collect emotional annotations, which are later associated with collected signals and used as ground truth. Depending on the study design, one can choose from different labeling strategies, e.g., labeling physiology with stimuli labels is easy in laboratory studies but impossible in field studies. Methods for labeling are detailed in Tab. 3.8 and in Tab. A.1 (*Ground truth* column), with methods for lab studies depicted in Fig. 3.4, and for field in Fig. 3.5.

The aforementioned labeling using stimulus type requires relatively little effort, but inductions have to be assigned with expected emotion. It can be done either by experts or by regular people. Twelve of the reviewed papers utilized such a method (Tab. 3.8), with six of them using additional self-assessment for validating

experienced emotions. Rattanyu et al. [95, 96] discarded samples where preassigned labels and self-reported emotions did not match.

Another way of assessing emotions externally, is to employ experts, e.g., psychologists. These experts can recognize emotional states based on subjects' facial expression [102]. Other works decided to utilize emotion recognition systems [63, 101, 107]. Dao et al. [63] trained their own model using self-assessments collected during the first part of the study and allowed participants to validate these predictions.

The most popular way of obtaining emotional labels was through self-assessments. Using this method of labeling is natural in field studies, where it is virtually impossible to find another source of labels, but it was also popular in lab studies. In total, it was used in 23 papers (68%, eight field studies, Tab. 3.8) and served as the source of labels in 17 of them.

In the laboratory setup, subjects filled out self-assessments either on a computer, a mobile device, or on paper (Tab. 3.9). We deduced from the context that Lisetti et al. [89, 90] and Zhao et al. [79] used paper questionnaires in their experiments, as the information was not provided clearly. Similarly, for two other articles [80, 101], we deduced from the experimental setup that authors utilized mobile devices, like phones or tablets, to collect questionnaire data.

As many as 72% of papers utilizing laboratory scenarios presented questionnaires right after the emotion induction. Another approach was shown by Martens et al. [85]. Their participants participated in an hour-long study session, where annotating once at the end would be impractical, so researchers decided to interrupt the stimulation every 270 seconds with a questionnaire. In laboratory studies, as researchers know the types of stimuli, answers to questionnaires were used either to directly label physiological signals (10 papers, Tab. 3.8) or to validate if targeted emotions were successfully induced (6 papers).

In several papers, the authors failed to explain the labeling procedure clearly. Setiawan et al. [108] did not describe emotion labeling. Hu et al. [88] provided no such information in their paper, but the content suggests that emotions were labeled either based on an expert's assessment or by the subjects themselves. Annotation procedure and ground truth source are similarly unclear in Nasoz et al. [87] paper. They used self-assessment questionnaires during the study but provided no information regarding physiological signal labeling.

Table 3.8: Methods of emotion labeling (ground truth); '?' – deduced by us (from [16]).

Method	Used by
Labels assigned to stimuli	Expert-annotated: [89–91] Crowd-annotated:[94]?, [80, 89, 90, 92, 93]
Labels assigned by system	[63, 101]
Labels assigned by experts	[102]
Stimuli labels validated with self-assessment	Expert-annotated: [97, 98] Crowd-annotated: [95, 96, 99, 100]
Self-assessment	[107]?, [11, 63, 76, 78–86, 103–106]
No info	[87, 88, 108]

Labeling in the lab

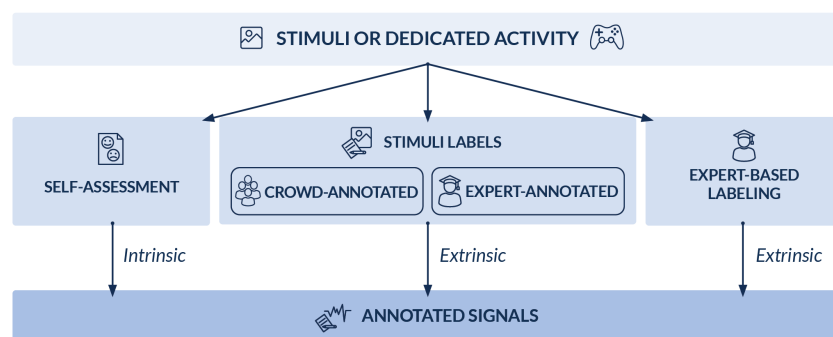


Figure 3.4: Possible ways of labeling physiological signals with emotions in the lab studies (from [16]).

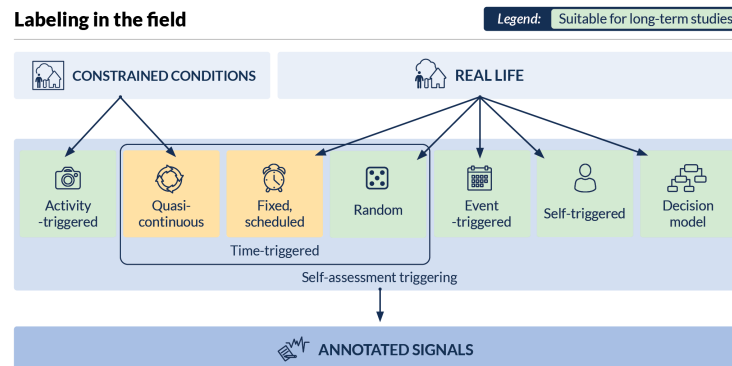


Figure 3.5: Possible ways of triggering self-assessment in the field studies (from [16]).

Table 3.9: Possible ways of collecting self-assessment (from [16]).

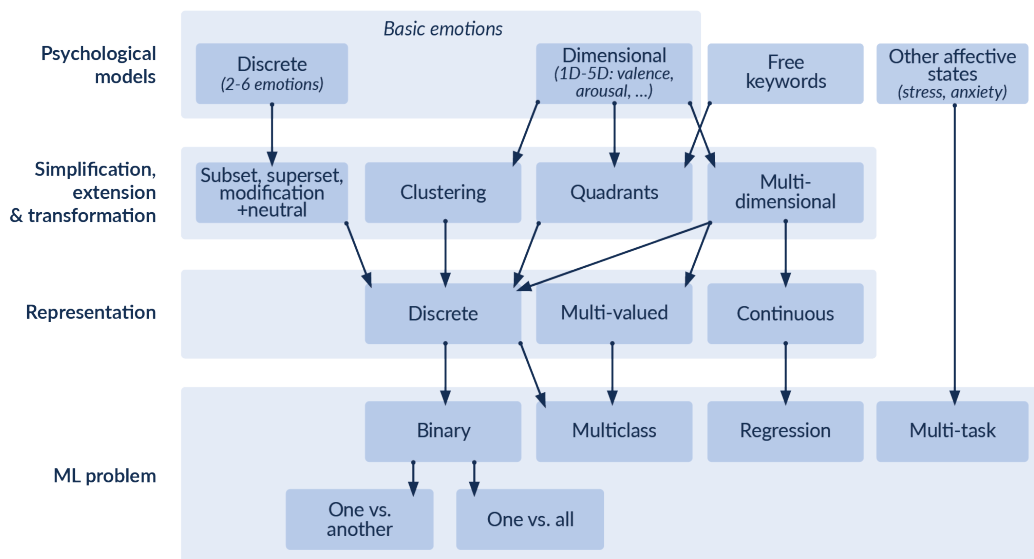
Environment	Tool	Used by
Lab	Paper	[89]?, [90]?, [79]?
	Computer	[95, 96]
	Mobile device	Tablet: [84]
		Phone: [85]
	No info	Not specified: [80]?, [101]?
Constrained field	No info	[78, 81–83, 87, 97, 98]
	Mobile device	Phone: [63, 103–105]
Field	No info	[107]
	Mobile device	Phone: [11, 76, 106] Watch: [76]

3.8 From theoretical model to reasoning task

Discrete psychological models often provide sophisticated dependencies between emotions, e.g., the position of the emotional states relative to other emotions in Russel’s Circumplex model, or Plutchik’s wheel of emotions, e.g., *joy* and *sadness* are opposite to one another, and *anger* is equally close to *disgust* and *anticipation*. When translated to simple binary or multiclass classification problem, all of those psychological relations are lost, and emotion categories become just an unordered set of distinct elements. Likewise, multidimensional models also have to be converted to discrete values, e.g., by grouping them in quadrants later treated as independent classes, and leading to a 4-class classification problem [79, 81, 91] (Fig. 3.6).

On the other hand, in some cases, authors respected the orthogonality of emotional dimensions and predicted each of them using a separate predictive model. It resulted in either two binary models [79, 82, 83, 108], one for each dimension; or two 3-class models [11, 80]. Another approach to addressing multiple discrete emotions was to create binary machine learning models for a given emotion in a one-against-all setup, e.g., a model predicting $1=\{sadness\}$ vs $0=\{anger, fear, happiness, relaxed\}$ [98]. Other methods included predicting one category against another for every combination of emotions. This approach was regardless of the initial models of emotion, discrete [102] or dimensional [92]. Lastly, Schmidt et al. [11] gathered self-assessments for *arousal*, *valence*, *stress* and *anxiety* and trained multi-task models with separate classification heads for each problem within one

(A) Transition of psychological models to ML problems



(B) Quadrants

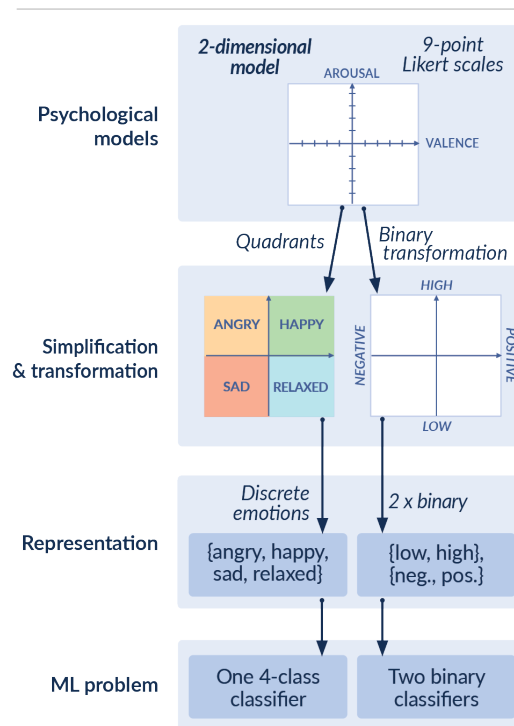


Figure 3.6: (A) Transition of original psychological emotional models into machine learning used in experiments; and (B) the 2-dimensional emotional space *arousal-valence* converted to four quadrants, next exploited as (1) 4-class classification or (2) two binary models, one for each dimension (from [16]).

machine learning model.

Regression was rarely considered in reviewed studies, with only three papers predicting continuous values [85, 86, 105] and one article using regression to measure the effect that affect has on productivity [107].

All in all, initial models of emotion were usually converted into simple machine learning problems, either discrete or binary. These simplifications may be caused by the relatively small sizes of psychophysiology datasets, making harder tasks, such as multiclass classification with a big number of dimensions, rather unfeasible.

3.9 *Learning case definition*

We define the *learning case* as the basic unit of knowledge provided to a machine learning model. In lab studies, knowledge about the study procedure can be used to, e.g., use signals collected during a whole emotion induction [97], its part [79], or combine signals from several stimuli of one type [93] (we assume using the whole recording, as signal splitting was not mentioned). In field studies signals are collected without break for a long periods of time. Therefore, dividing them into fragments is necessary, but extracting proper learning cases is challenging, as exact time of stimulation is at best uncertain, or even unknown.

In the emotion recognition field, dividing collected signals into segments is often called *windowing*. It is often used to extract many many learning cases from one registered signal to increase number of possible learning cases. Windows are often all labeled with the same label as the original signal [99, 100]. This method may be especially useful when working with deep learning models, as in general they require large datasets for training, e.g., Schmidt et al. [11] extracted up to 240 cases for each labeled affective state. Divided signals may also be treated as a time series and fed as such into a neural network with recurrent layer [104].

Papers that used windowing (41%) also considered different aspects of the process. In five of them [85, 92, 96–98], authors utilized windows spanning the whole recorded signal. In papers that divided recordings, window sizes varied from one second [81] up to 180s or more [92, 106]. Additionally, used windows could be adjacent [79, 104], or overlapping [11, 81, 83, 99, 100, 105]. Also, it is possible to use different window sizes depending on the type of signal, e.g., Wampfler et al. [84] extracted features from two sources using a ten-second-long window for physiological signals and the whole stimulus length for data from a stylus.

3.10 Machine learning models

Usually, in what we call a *classical feature-based* approach, the collected raw signals must be preprocessed and synchronized before training emotion recognition models, and descriptive features must be derived. After that, researchers train and optimize simple classifiers or deep neural networks. Extracting meaningful and informative features requires expert, domain-specific knowledge about utilized sensors and collected signals. On the other hand, in an *end-to-end* approach, acquired signals are directly passed to the deep learning architectures. It requires careful consideration of model design and a training procedure to ensure that the essential patterns will be found automatically. The end-to-end approach is a promising direction [11, 121, 122]; however, in the reviewed papers, the authors rarely applied it.

In Tab. 3.10, we summarize the approaches used at different stages of the ML model training process. Utilizing classifiers, such as decision tree (DT), k-nearest neighbors algorithm (KNN), support vector machine (SVM), or simple Neural Networks to solve a multiclass problem was the most popular among reviewed articles (88% of papers). In total, only 21% of studies applied deep learning algorithms, with only four papers using more sophisticated deep neural network architectures (convolutional neural networks, CNN; or long short-term memory, LSTM). Multilayer perceptrons (MLP) were utilized in the earliest deep learning experiments [89, 90] but remained used in later years. Some authors used MLP in later years despite also experimenting with more advanced architectures like convolutional (CNN) or long short-term memory (LSTM) networks [81, 104]. Moreover, Saxena et al. [86] used MLP to address a regression task. Other approaches were presented by (1) Schmidt et al. [11] that experimented with training CNNs in the end-to-end fashion, and (2) Tizzano et al. [100] that trained an LSTM model using transfer learning.

Although most of included papers (56%) analyzed performance of multiple classifiers or deep learning architectures [11, 76, 79, 81, 83–87, 89, 90, 92, 93, 99, 100, 102, 103, 106, 108], only Kanjo et al. [103] and Schmidt et al. [11] compared different ML training methods – feature-based approach vs. end-to-end deep learning. In both cases, deep learning provided better outcomes, but the authors did not provide statistical tests to confirm this.

Out of assumed machine learning tasks, the multiclass classification was the most common approach (74% of papers, Tab. 3.10). In eight papers, the ML problem was defined as binary classification, often reducing the complexity of the initial model of emotions assumed in the study (Sec. 3.8). Only three papers defined the problem as

a regression [85, 86, 105], and only Schmidt et al. [11] defined a multi-task problem, which they later classified using a single ML model giving multiple classification outputs. Comparisons of different ML problem types, e.g., binary classification and multiclass classification separately [11, 79, 80, 84, 92, 99–102] were performed in nine papers (26%).

3.10.1 Analyzed approaches

When searching for the best-performing approach, one has to consider many factors other than ML models and tasks, among others, the number of windows originating from one learning case [83], the or amount of overlap between windows and their sizes [83, 104, 106]. Tizzano et al. [100] investigated the feasibility of transfer learning for creating the personal model but failed to achieve any valuable conclusions.

Other factors analyzed in reviewed papers included differences between devices [82] and signals (e.g., physiological vs. environmental) [76, 84, 103, 104]. Setiawan et al. [93] showed that modality fusion at the feature level performed better than the one at the decision level. Similarly, other authors studied impact of different feature sets [86, 99, 102, 106], cardinalities of these sets [92, 95, 96, 98], or methods for feature selection [79]. Feng et al. [102] focused on feature generation for time-frequency analysis of the EDA signal, and after analyzing four mother (prototype) wavelet shapes, they concluded that Complex Morlet was the best.

Also, experimental setups and the impact of ground truth choice can be analyzed. Albraikan et al. [80] compared results from using ground truth of different origins, i.e., annotations with self-assessment vs. expected emotion, and achieved better model accuracy using the latter. They suggested that annotations provided by subjects may be biased and that stimuli types are more objective as ground truth, but they did not report any statistical analysis. Quiroz et al. [99] designed multiple experimental setups: (1) walking after watching a movie; (2) walking after listening to music; and (3) walking while listening to music, classified emotions in those setups, and achieved the best results in the third setup. Two other papers utilized multiple datasets when performing emotion recognition [80, 83].

3.10.2 Hyperparameter tuning

In machine learning, hyperparameters include different attributes or methods used when training a model, e.g., learning rate or activation function in neural network training or a number of trees in a Random Forest. Hyperparameter tuning can be

used to find a combination of parameters that result in the best-performing model. It can be done manually (e.g., grid search), semi-automatically (e.g., Bayesian optimization), or by using fully automated tools like auto-sklearn [123].

Hyperparameter testing is used to find the best parameters for the model while maintaining its generalizability and avoiding overfitting. However, model optimization introduces bias for the dataset used, often a validation set. Thus, to reliably measure model performance on unseen data, such an optimized model should always undergo final verification on a previously unseen test set. Also, optimized performance metrics should be chosen with the problem and data profile in mind, e.g., data imbalance.

Only a few of the reviewed papers utilized hyperparameter optimization (Tab. A.3). Some authors conducted simple model tuning, i.e., adjusting the number of trees in Random Forest [83], analyzing different SVM kernels [98, 102], or performing randomized parameter search [84]. Only Nakisa et al. [81] compared several optimization algorithms and selected an approach based on optimization results.

3.11 Model validation

To obtain a machine learning model that is ready for real-world inference, one needs to perform thorough testing and optimization. They are crucial to avoid situations where models trained on data from one source, e.g., own laboratory dataset, fail to generalize to data from another source, e.g., real-life signals. This testing should be performed in relation to the assumed emotional ground truth (reported emotions), relations that are present in the data, and a final goal.

3.11.1 Quality measures

Training predictive models include optimization of a chosen quality measure. This measure not only impacts the model but is often used to evaluate the model's performance. Most of the reviewed papers (82%, Tab. 3.10) utilized *accuracy* measure defined as a ratio of correct predictions (consistent with the ground truth) to all model predictions (sum of correct and incorrect). Accuracy is easy to understand and interpret but also ignores imbalance in the data, e.g., a model predicting only majority class on a dataset consisting of 90 samples of anger and 10 cases of joy achieves 90% accuracy. Such a model may be mistakenly seen as well-performing when looking at the metric alone, even though without the ability to discern between different emotional states, it would be useless in practice.

Six papers (18%) [76, 78, 80, 81, 88, 102] used only accuracy despite having unequally distributed classes A.3. Four papers (12%) utilized quality measures that, to some degree, account for the distribution of classes: macro F-measure [11, 83], macro-avg AUC [84], or confusion matrix [104]. Only three (27%) out of 11 papers using imbalanced datasets considered implementing training techniques for imbalanced data, such as balanced class weights [84], equal size sampling [98], or binning values of Likert scales (adjusted ranges) [11]. Finally, only eight papers (24%) validated their results using statistical tests [11, 81, 83, 85, 95, 96, 101, 105].

3.11.2 Validation procedures

We can group validation scenarios used in emotion recognition studies into six approaches, based on the focus on particular dataset components over which the split into train and test examples is performed (Tab. 3.11):

1. **Non-specialized validation**, which is a standard method used in machine learning. Splitting is done over the entire dataset without accounting for subject- or stimulus-specific properties. It can take form of (a) simple train-test split [86, 95, 96, 103, 106], (b) stratified split relative to the output class distribution [98], or (c) k-fold cross-validation [89, 90, 92, 108] (Fig. 3.7A).
2. **Intra(within)-subject validation**, where the data of a particular subject is split into train and test sets. In this approach, ML models are trained on the part of subjects' cases and tested on the rest of the data. Such validation is needed when researchers assume the necessity of having subject-related data for training so it does not measure the model's reasoning abilities toward new subjects. It can be performed as (a) simple one-time split [88, 104]; (b) within-subject cross-validation [91, 99]; (c) leave-k-cases-out for each subject [85]; or (d) leave-target-questionnaires-out (LTQO) [11].
3. **Inter(between)-subject validation** is an approach specific for modeling human-related data, as it emphasizes the need for user-independent validation and assigns all data of a particular subject to either train or test set. This approach tests the model's ability to generalize for unknown people, and can be done as a one-time split [97], or cross-validation over subjects [11, 79–81, 83–85, 91, 93, 99, 102] (Fig. 3.7C).
4. **Inter-intra-subject validation** suitable for finding the model's capabilities for generalized predictions and its potential for further personalized adjustment. In

the studied literature, it was done as cross-validation (one subject at a time, LOSO) combined with the repeated random split on the test subject's data [100].

5. **Between-stimuli validation.** Models are trained on data from a set of stimuli and tested on another set of stimuli. It can take the form of a split based on stimuli (a) such as videos [83] (Fig. 3.7B), or performed activities [84] (Fig. 3.7D).
6. **Across time validation** test models' ability to generalize between periods in time. Training is performed on examples from earlier periods, and testing on samples collected later in time [63, 76] (Fig. 3.7E).

The choice of the validation method affects the machine learning model's qualities that are being tested, e.g., its generalizability relative to (a) subjects in between-subject validation or (b) stimuli in between-stimuli approach (Tab. 3.11). Depending on the experiment setup, domain-specific validation approaches may be better suited than classical cross-validation, thanks to their context-respecting properties [124], e.g., training on a subset of stimuli and testing on the rest [83] (between-stimuli validation). Moreover, researchers should consider if, in their setup, it is better to train general or personalized classifiers - a question considered only in two studies [100, 104].

Drawing conclusions in emotion recognition studies requires careful consideration of different relationships present in the data. Usually, handling inter-subject variability proves difficult, with some researchers abandoning between-subject validation due to the subpar performance of their models [83]. Overall, this validation scheme appears to best reflect a scenario where models should perform well on previously unseen people, therefore reflecting a cold-start problem present in the real world.

Combining between-subject and within-subject validation is an interesting approach, which allows utilizing general models trained on data from known subjects (between-subject validation) and adjusting these models training for previously unseen people, for example, by training new layers on data from the subject excluded from training (within-subject validation). In this case, the final testing is performed on the remaining data of the test subject. The remaining subject's data can be split using the monte-carlo approach [100] or k-fold cross-validation (not considered in any paper), or across-time validation (also not considered).

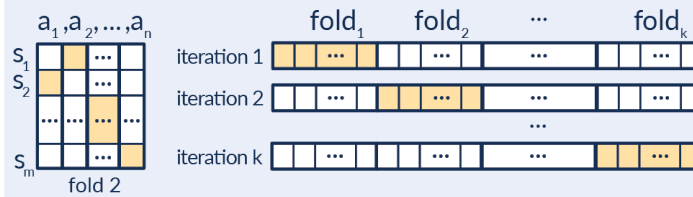
A proper choice of validation procedure leads to good insights into the trained model's properties and its generalization or personalization ability. It can be mea-

Validation

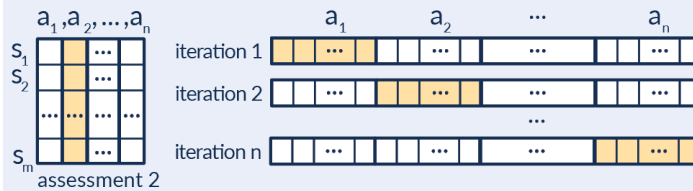
Legend: Train set ... Test set ...

Cross-validation

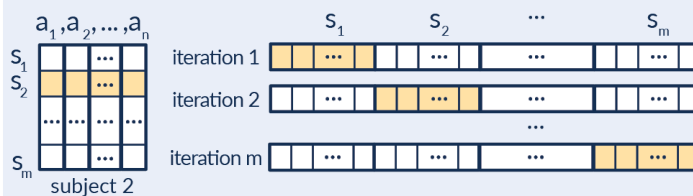
(A) Classic k-fold cross-validation over all cases



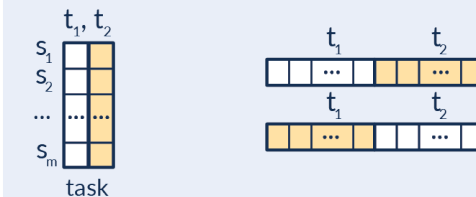
(B) Intra-subject: leave one assessment-out, e.g. Leave-One-Video-Out (LOVO)



(C) Inter-subject: Leave-One-Subject-Out (LOSO)



(D) Task-based validation



Across Time Validation

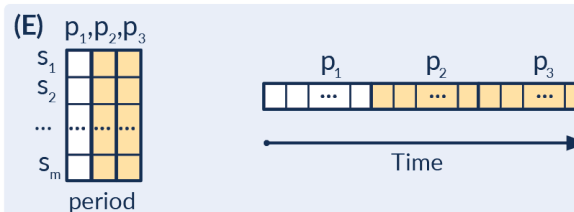


Figure 3.7: Selected validation methods used in emotion recognition; a_i denotes the i th assessment, s_j – the j th subject. The matrix may not be fully filled out, since some examples may be removed or uncollected (from [16]).

sured, e.g., using standard deviation calculated over cross-validation folds. In the case of between-subject validation, e.g., leave-one-subject-out, it provides a margin of the model's ability to predict a previously unseen person's emotions. Unfortunately, not many articles report information about the generalization performance of their methods. For example, 5 out of 12 papers that utilized between-subject cross-validation failed to report any information related to differences between data splits, e.g., individual error values, mean absolute error, or standard deviation [79, 80, 93, 100, 102]. Additionally, authors commonly failed to consider the data imbalance (88% of papers) while applying imbalance-sensitive quality measures, such as accuracy (82% of papers). Furthermore, a minority of articles (24%) investigated their results using statistical tests.

Four papers went beyond using a single validation procedure, and analyzed their models in multiple setups: LOSO vs. intra-subject 10-fold cross-validation [91, 99, 100] and LOSO vs. LTQO [11]. In all of those, validation setups where models could learn from data of specific subjects resulted in higher F-measure values than the subject agnostic validation (LOSO).

3.12 Discussion and identified challenges

In this work, we investigated the possibility of utilizing signals collected with wearables to recognize real-life emotions. During the review, we realized that most of the authors applied similar procedures in their studies, i.e., they recruited participants, prepared stimuli and annotation methods, collected signals and processed them while also collecting self-assessments, and finally, trained and validated machine learning models. Moreover, we identified seven scenarios of emotion recognition and how researchers converted psychological models to machine learning tasks.

3.12.1 Study Design

No article reviewed in SLR was exhaustive with respect to all research components. In particular, (1) only five studies (15%) considered more than four emotions (five/six discrete ones or six emotional regions); (2) in nine papers (26%), authors obtained ethical committee approval; (3) in 14 papers (41%) written consent was collected from participants; (4) in ten papers (29%), context data was considered, e.g., alcohol consumption, physical exercises; (5) in 14 papers (41%), subjects' health conditions were checked, e.g., personality disorders or cardiovascular diseases; (6) in three papers (9%), the labeling procedure was not described at all, and in seven papers (21%), inductions were not verified by self-assessment; (7) out of 11 papers using

imbalanced data, only three treated it utilizing appropriate techniques; (8) only seven studies (21%), performed any adjustment to models' parameters, and only three optimized them; (9) inter-subject validation procedure (subject-independent validation) was used in only 12 papers (35%); (10) in only nine papers (26%), results were tested for their statistical significance; (11) no authors using their own dataset shared it; (12) only one paper had an accompanying source code published. Additionally, most studies are relatively small in scale, with only four papers (12%) conducting research on groups of over 50 subjects and none investigating more than 100 people.

Most of reviewed studies were carried out in laboratory conditions. Only eight papers considered real-life environment with its unique challenges (Tab. A.2). In particular, data collection in the field has to employ convenient and non-annoying labeling methods. Additionally, as some emotions are rare in everyday life (e.g., fear or disgust), researchers can expect a high imbalance of classes in collected data.

3.12.2 Emotional models and ML problems

The reviewed articles extensively utilized simple models of affect, such as *low arousal-high arousal* or *no stress-low stress-high stress*. It was mainly driven by strong correlations between arousal or stress and some biological signals – BVP or EDA [12, 125]. However, emotions are complex states with a multidimensional nature. Respecting all aspects of emotion still remains a challenge. Most of the studies considered up to six basic emotions (Tab. A.1), despite the fact that researchers postulate much more extensive models of emotion. For example, Du et al. [126] identified 21 states from facial expressions, while Cowen and Keltner [127] observed 27 distinct categories in self-reports. Nevertheless, distinguishing a large number of categories would also make the recognition problem much more difficult.

The assumed model of emotions directly impacts the design of the detection model. We would expect to treat multidimensional models as a multi-label classification problem. However, a large number of training cases would be required to train such a classifier, and none of the papers approached reasoning in such a way. It was only partially addressed in seven papers utilizing dimensional emotional models and multiple independent regressors [85] or classifiers [11, 79, 82, 83, 92, 108].

Another challenge lies in creating models combining general and personal knowledge [100]. Such models (preferably deep learning architectures) would need to learn from the whole population of subjects and from the personal signals of a

targeted individual. Moreover, the specificity of human physiology and subjectivity of affect require careful consideration when designing solutions and validation procedures. For example, subject-dependent (i.e., non-inter-subject) validation usually results in relatively high metric values because data from the same subjects are used for both training and testing [100].

3.12.3 Data collecting

Collecting data is challenging, especially when done in real life. When collecting large amounts of data for training machine learning models (especially the deep learning ones), using unobtrusive multipurpose devices and novel triggering methods [15, 128] may prove useful.

We observed that most studies utilized similar hardware devices (Tab. 3.6). It may be caused by the fact that provided signals differ between off-the-shelf wearables, and in most of them, getting raw physiological recordings is impossible. Additionally, most wearable sensors provide signals sampled with low frequency and quality, causing their quality to be much worse than respective medical-level devices, especially when recording in motion [117].

In the review, we focused on studies potentially applicable in the field, so they had to utilize portable wireless wearables. Apart from those, other complementary signals and data may be collected in real life, e.g., our activity [129], voice [130], or face recording [131] (if smartwatch has a built-in camera), and potentially improve emotion recognition accuracy. However, each new device adds to the complexity of the data synchronization process. Moreover, collecting some modalities in real life may be challenging, e.g., facial expressions when a person moves or their voice in a loud environment. Also, continuous monitoring consumes much energy. Thus, it may be necessary to charge devices frequently [128].

3.12.4 Data processing

Typically, physiological signals recorded with wearable sensors contain a lot of artifacts that mask affective information. It necessitates using various methods to reduce distortions, e.g., filtration, winsorization, interpolation, or their various combinations.

Feature extraction methods in analyzed articles utilize only a small part of all available approaches present in the wider literature. For example, out of different signal decomposition methods, only wavelet transform was used, ignoring methods

such as empirical mode decomposition (EMD), Hilbert-Huang transform (HHT), matching pursuit algorithm (MPA), or variational mode decomposition (VMD).

Surprisingly, no article performed feature selection, although it may reduce machine learning models' complexity and increase their accuracy. In future studies, authors could consider methods such as relief algorithms, analysis of variance (ANOVA), or minimal redundancy maximal relevance procedure (MRMR). Also, as it is still debated how long emotions last, the proper length of signal windows remains undefined, and authors have to rely on their own experience when choosing window size.

3.12.5 Machine learning

The majority of studies (76%) utilize simple machine learning methods based on decision trees or hyperplanes (for instance, SVM, LDA), which can achieve good quality when solving simple problems. However, several works showed that more complex deep learning architectures better model complex relations, especially in multimodal emotional data [11, 81, 100, 104]. Overall, deep learning models, especially end-to-end ones [11, 104, 121] were rarely employed despite their great potential.

Many papers failed to address imbalance in learning samples, even when it was apparent [76, 78, 80, 81, 88, 102]. Moreover, they often utilized accuracy measure (fraction of correct classifications) as the only metric to assess classification algorithm quality, leading to (most probably) overestimated results.

Furthermore, model efficiency is often evaluated in terms of quality measures. While practical, numerical measures only provide information about predictive abilities on test data, which can be heavily dependent on the particular dataset or validation strategy. Such a measure does not assess model performance in real life, where some necessary conditions may be violated, or even new or co-occurring emotions may be present, e.g., fear and anger after dropping a smartphone into water. In such a case, a model trained on well-discernible emotions may not be able to determine the correct class (emotion) accurately or will wrongly suggest the most probable emotion.

3.12.6 Research replicability and comparability

It is well-known that many issues with commensurability can be found in research utilizing artificial intelligence. It is also true for the reviewed articles because:

1. authors utilize different emotional models (discrete emotions or emotional dimensions) that are not equivalent, e.g., Hu et al. [88] and Pollreisz et al. [78] utilized four discrete emotions, but one included fear [88] in their model, while the other used pain [78] (Sec. 3.6)
2. same emotional models were transformed to different machine learning problems (Sec. 3.8)
3. different kinds of data were used in experiments – researchers utilized various modalities, gathered using different hardware differing quality and sampling frequency (Sec. 3.5)
4. used quality measures and validation procedures are incomparable (Sec. 3.11)
5. utilized data and code are often not available for other researchers – only Romeo et al. [83] utilized publicly available dataset (lab, DEAP [132]) and published their code

To improve commensurability of affective research, authors could use in their publications (1) well-established emotional models without unnecessary modifications (e.g., Ekman-Friesen [39], Plutchik [133], or dimensional arousal-valence), (2) popular validation procedures (e.g., LOSO), and (3) suitable and widely used quality measures (e.g., F-measure, ROC AUC, accuracy). We advocate for including already established methods and models in conducted experiments instead of only focusing on own designs and experiments. Also, to improve replicability, authors should publish their data and code whenever possible.

3.12.7 New arising challenges

Emotions in real life are usually complex, and people may experience different combinations of basic emotions at the same time [86]. In an uncontrolled environment, they are impacted by the context that a person experiences at the time [134] (e.g., a specific situation, such as a job interview). Therefore, questionnaires for in-the-field studies should not rely on predefined emotions but rather try to gather more general affect measures and accompanying contextual metadata, ideally using free-text responses. In the research that considers distinct emotional states (categories), they should be treated as multi-label classification or multivariate regression [86] and return estimates for all possible states at the same time, e.g., separate probabilities for all emotional states, because in real life multiple of them may occur at the same

time. Also, additional personal context, such as knowledge about other affective states (e.g., stress, anxiety, or mood) or personal traits (e.g., personality), may help in discovering meaningful patterns and relationships in data, therefore improving prediction quality. It may be realized using, e.g., multi-task deep learning models [11, 135].

Also, novel deep learning methods may also be useful for signal processing signals and generating their representations. One of most promising methods is self-supervised learning, which allows using unlabeled data for model training models, as collecting emotion annotation is a long, difficult, and noisy process. Autoencoders could be used to impute missing measurements, therefore improving signal quality. Transformer networks could be used to better tackle multimodal data and improve reasoning about emotions. Also, new developments in AI explainability may allow us to better research the nature of emotions by identifying specific links and patterns between emotions and input data.

We believe that personalization of reasoning is necessary, especially the development of methods that would respect differences between people in their physiology and subjectivity of emotional experience. Also, combining knowledge about individual subjects (personal) with population-wise (general) patterns [100] seems promising. Furthermore, as time impacts people's behavior, physiology, and perception, researching methods that would allow updating models using incoming data from an individual and a general population, e.g., incremental and active learning methods. Also, together with the development of new methods for emotion modeling, new validation procedures also have to be researched, as such models require procedures for testing both their generalization and personalization abilities, e.g., new inter-intra-subject methods.

Due to many differences between laboratory and real-life conditions, reasoning in everyday life using models created with data from laboratory studies will probably show poor accuracy. Therefore, researchers should focus their efforts on collecting datasets with annotated emotional events from real life. However, it requires precise, reliable, and unobtrusive sensing devices suited for everyday use. Ideally, they would allow collecting multimodal data, e.g., BVP together with ECG, EDA, acceleration, or speech data, all of excellent quality and high sampling frequency. Although measuring devices have become more unobtrusive and provide signals of better quality than in the past, they have yet to reach the quality resembling that of their laboratory counterparts.

Lastly, we believe that for affective research to advance, the community has to focus on conducting large-scale studies in real life (with many participants measured over long periods of time) and promoting open science. Recorded data should contain contextual information so more factors can be included in analyses, such as physical activity (e.g., hiking vs sitting on a couch) or other temporal context. Such studies should be designed in multidisciplinary teams, with experts from a wide range of research fields involved. Also, monopolizing collected datasets and not publishing the code for developed solutions hinders progress in affective computing. Such practices not only slow down progress due to fewer resources available for researchers but also make developed solutions impossible to recreate and verify.

We are not alone in our findings, as other researchers provided similar observations. Larradet et al. [136] commented that the majority of studies are conducted in laboratory conditions and that utilizing such data may cause issues when creating reasoning models for real life. They also created a method for assessing if emotional datasets can be used in developing methods for in-the-field reasoning about emotions, and most of the examined datasets performed poorly (their charts occupied a relatively small area).

Table 3.10: Approaches, methods, and measures used at particular machine learning stage. '?' means it was not described but inferred by us (from [16]).

Stage	Approach	Used by
Classification type	multiclass	[11, 76, 78–81, 84, 86–90, 92, 94, 97, 99–104], [95]?, [96]?, [106]?, [63]?, [91]?, [93]?, [107]?
	binary regression	[79, 82, 83, 98–100, 102], [108]?, [85, 86, 105]
ML models	classical (KNN, SVM, etc.)	[11, 76, 78, 79, 81–84, 86–100, 102, 103, 106, 108]
	deep (CNN, LSTM, etc.)	[11, 81, 100, 104]
Quality measures	accuracy	[76, 78, 80–84, 86–104, 106, 108]
	F-measure	[11, 80, 83, 99, 103, 104]
	Other	[78] – conf. level; [102] – AUC, precision, recall; [79] – correct classif. ratio; [104] – precision, recall, error rate, RMSE, confusion matrix; [83] – ROC curves, confusion matrices; [84] – micro-avg AUC, macro-avg AUC; [85] – MAE, RMSE, Pearson correlation coefficient
Imbalance in learning samples	not considered	[63, 76, 78–83, 85–97, 99–108]
	considered	[98] – equal size sampling; [84] – RF with balanced class weights, macro-avg AUC; [11] – converting Likert scales into bins (adjustment of ranges)
	balanced data	[86, 89, 90, 92, 97, 99, 100, 106], [87]?, [80]?, [91]?, [94]?
	imbalanced data	[11, 76, 78, 80, 83, 84, 88, 98, 102, 107], [81]?
Statistical tests on results	none	[63, 76, 78–80, 82, 84, 86–94, 97, 98, 100, 102–104, 106–108]
	applied	[95, 96] – ANOVA, LSD; [105] – p-value, analysis of beta coefficient; [99] – p-value; [101] – McNemar’s test (within- subjects chi-squared test); [81] – ANOVA; [83] – Wilcoxon’s signed-rank; [11] – Pearson correlation coefficient; [85] – student’s t-tests

Table 3.11: Reasoning model validation methods and strategies for dividing data into training and test sets (from [16]). The number of '-' or '+' denotes the ability to estimate generalization level of the considered models. '?' means we had to deduce it.

Subject to split	Validation type	Generalizability	References and details
Whole set	Validation on the whole set	- - - -	[101]
	One-time split over all cases	- - -	train and test [86, 95, 96, 103, 106]; train, validation and test [86]
	Stratified one-time split over cases	- -	train and test [98]
	Cross-validation over cases	-	k-fold [82, 89, 90, 92, 108]; Monte Carlo [100]
Within-subject	One-time split over each subject's cases	- -	train and test [88, 104]
	Within-subject cross-validation	+	10-fold, independently for each subject [91, 99]; leave-one-out (LOO) – leave-one-observation-out of a given subject [85]
	Leave- k -assessment(stimuli)-out	+	Leave-One-Video-Out (LOVO), 10-fold cross validation over video [83]; leave-target-questionnaires-out (LTQO), a stratified N -fold split over classes and questionnaires (train/test/validation) [11]
Tasks	Between the different tasks performed	++	train and test [84]
Time	Across time validation	++	train and test [63, 76]
Between-subject	One-time split over subjects	++	train and test [97]
	Leave-one(k)-subject-out (LOSO) cross validation	+++	[11, 80, 81, 83–85, 91, 93, 99, 100], [79]?, [102]?;
Between-within-subject	LOSO + intra-subject repeated random split	+++	[100]
No info	No info	N/A	[63, 78, 87, 94, 105, 107]

Chapter 4

PERSONALIZED EMOTION RECOGNITION REVIEW

Calls for respecting individual differences between people when working on emotion recognition are as old as affective computing itself. In one of her early works, Picard [10] stressed the importance of within-person analyses and deemed universal emotion recognition unnecessary. She reasoned that not only do people differ in their perception and expressions of emotions, making training universal models problematic, but also that personal ER may be preferred by people who would like to share their emotions only with trusted devices and people. To this day, researchers tend to explore individual differences between people [137]. It seems especially important for emotion recognition from physiological signals, as inter-subject variability occurs in both physiology and affect measures [11].

Researchers understand personalization of affect recognition in various ways, e.g., Tizzano et al. [100] trained separate model for each subject, whereas Yu et al. [138] introduced a personalized component to an already trained general model. In this chapter, we describe and categorize approaches to a personalization of emotion recognition. Moreover, we provide a short analysis of signals, ML tasks and algorithms, and validation approaches found in the literature.

While machine learning algorithms require large amounts of data to capture general patterns and dependencies, personalization may benefit researchers working on a small sample size, as it focuses on capturing person-specific information in place of general patterns. Most of the reviewed articles performed experiments on relatively small datasets of up to 40 subjects. The biggest dataset utilized for experiments with personalization was data from the SNAPSHOT study [139], which contains data from about 255 participants but remains unpublished.

We decided to include literature that considers not only emotions but also affective states such as mood or stress in this chapter. At the time of the search, we could find only a few articles focusing on personalized emotion recognition. Thus, we found it necessary to broaden the scope of the review to cover a larger number of methods used in personalized affective computing.

Table 4.1: Personalization approaches in reviewed studies. Legend: ? - deduced by us.

Personalization subject	Subject aggregation	Personalized procedure	Personalized model	Used by	
Individuals	-	New model for each subject	-	[61, 76, 99, 100, 140–152], [153]?	
		Transfer learning	Personalized layers	[138]	
			-	[64, 100]	
		Multitask training	Personalized layers	[62, 154–156]	
Groups of people	-	New model for each group	-	[138, 144, 152, 157, 158]; [62, 159]?	
		Questionnaires	New model for each group	-	[154, 155, 160–162];
		Multitask training	Personalized layers or parameters	[154–156, 160]	
	Tasks	New model for each group?	-	[151]	
	Random	-	Transfer learning	-	[64, 65]
			Retraining	-	[65]
			Physiology	New model for each group	-
Hybrid approach	-	New model for each subject	Personalized parameters	[164]	
		Mapping input into general domain	-	[165]	
		Physiology	Weighing samples	Personalized parameters	[150]
		Adding samples	-	[166]	

4.1 Personalization strategies

In general, personalization can be understood as creating a machine learning model for a specific domain. It can be achieved in many ways, e.g., by training domain-specific models or by adapting existing models to a domain of interest. The very nature of this problem is not specific to affective computing, and researchers from other fields (e.g., natural language processing [167] or computer vision [168]) focus their efforts on inventing and improving model training strategies. Even a reader who is new to the subject may be already familiar with one of the most well-known methods for adjusting the model – transfer learning [169, 170].

However, affective computing is strictly focused on people’s affective states and always utilizes human-related features for reasoning. Therefore, in affective computing, we use the term personalization to describe a family of methods for targeting an individual or their traits. It can be done in different machine learning stages, from data processing to model adjusting.

4.1.1 Subject of personalization

The first dimension of personalization is the choice of a subject. When creating personalized models, different people can be treated as distinct individuals or as members of a larger group who share similar traits. If we utilize group similarities or knowledge to improve the created model, the group essentially becomes a new subject, and we treat models as group-personalized. We treat the model as group-personalized even if people are simply treated as a part of a larger population with no regard to their similarities, as such models can still learn specifics of people included in the training.

Approaches to choosing personalization subjects differ between articles (Tab. 4.1). Although the most common approach was to treat participants individually, personalization performed on groups of people is also often used. Also, we classified three of the reviewed articles as a mix of individual and group approaches.

The most straightforward approach to group personalization considered all people as a part of a population, and their training and test datasets contained data from the same subjects. Similarly, Yu and Sano [65], and Wu et al. [64] focused on randomly created groups of participants. Although such approaches are not based on similarities between subjects, it may be possible for models (especially deep ones) to capture particular subjects’ specificity and use it for reasoning. The drawback, however, is that such a setup cannot be used for previously unseen people. This is

also the case in an article by Dessai and Virani [151], where authors focused on groups sharing an emotional experience (assuming a synchrony between them).

Knowing that personality may influence emotional experience, some authors decided to use it to aggregate subjects. In four papers [154–156, 160], authors clustered participants with k-means or k-prototypes algorithm based on their gender and answers to a Big Five Personality Model questionnaire [171]. Similarly, Tian et al. [161] used only the Big Five model together with the k-means algorithm to form groups. Can et al. [162] presented a slightly different approach and divided participants based on their Perceived Stress Scale (PSS) into three groups: low, medium, and high stress.

Similarities between subjects can also be found in their physiology. In four articles [142, 150, 163], researchers grouped people based on their physiological measures to create models for them. These similarities were also utilized in what we described as hybrid approaches, with Hernandez et al. [150] creating general models weighed using similarity between the general population and a person, or Bang et al. [166] enhancing expanding personal datasets with data from similar people.

4.1.2 Personalized procedures

The next stage that can be personalized is the procedure used to create the machine learning model. In reviewed articles, we identified seven different approaches to personalize training procedure: training a new model for each subject or group, transfer learning, retraining the model with personal data, multitask training, mapping personal data to a general domain, and weighing or adding training samples (Tab. 4.1).

The simplest strategy, i.e., training the machine learning model for each subject or group separately, was the most common approach. Although most researchers simply trained models on personal data, some utilized more sophisticated strategies. Exler et al. [76] considered not only separate training for each participant but also different types of decision trees. Lotfalinezhad and Maleki [148] and Muaremi et al. [143] performed separate feature selection for each participant before training personalized models. Hernandez et al. [150] trained models in two such settings: (1) a separate model for each subject, (2) weighing loss during the ML model training, and effectively creating a model for a group of similar participants.

Some authors utilized general models in their works, either in a transfer learning

or retraining scenario. In transfer learning, authors first trained general models and later fine-tuned them to fit target data better [64, 65, 100, 138]. Wu et al. [64] also explored another variant of this procedure, where they used a fine-tuned group model and performed further tuning to create individual models. On the other hand, Yu et al. [65] additionally considered retraining the deep learning model on the personalized data but did not provide much explanation of this approach. We deduce that they probably followed the same approach as in transfer learning and further trained the general model they already had.

Another type of personalized procedure was using multitask learning, where models were trained to learn reasoning for different people or groups simultaneously. It requires algorithms or architectures designed to provide multiple predictions (tasks) at the same time but allows common parts of models to gain more knowledge during training. In this review, four articles utilized such approach [62, 154–156].

Researchers who treated subjects in a hybrid manner, as a part of a population and as individuals at the same time, utilized different approaches than the ones described above. Shi et al. [164] trained general models with subject-specific parameters, and Hernandez et al. [150] created a general model but used a distribution of specific participant's labels to model parameters during inference, thus creating a personalized model using non-personal data. Bang et al. [166] created personal models, but if subjects did not have enough data, they used samples of a similar participant for training (either by expanding or replacing personal dataset), thus treating subjects as a part of a group. Unfortunately, there was no information on whether it was used for expanding only the training or both training and testing sets. Lastly, Gasparini et al. [165] mapped each participant's data to a general subject-normalized domain, common for everyone. They achieved it by resampling each person's data to match the desired heart rate frequency based on their heart rate during rest.

4.1.3 Personalized algorithms and deep architectures

Despite the recent popularity of deep learning models, classical machine learning still constitutes the majority of approaches in the literature on personalized affect recognition. Often, algorithms such as Random Forest (RF), Support Vector Machine (SVM), or simple neural networks (multilayer perceptrons, MLP) were used. Although classical approaches were most often trained on personal data, some authors decided to perform interesting modifications to standard algorithms.

Shi et al. [164] introduced person-specific parameters to SVM, which is responsible for personalized feature mapping. Rubin et al. [61] utilized an approach that differed from other classification approaches, as they performed the outlier detection. They fitted the Gaussian to the data of each subject and considered samples with low probability values as outliers (the threshold was fitted based on training data). Yu et al. [154] and Taylor et al. [160] utilized multitask learning (MTL) versions of classical machine learning models such as lasso regression, linear regression, support vector machines (SVM), or Bayesian hierarchical logistic regression.

In 17 papers, deep learning models were utilized, such as neural networks with convolutional (CNN, eight papers) or long short-term memory layers (LSTM, eight papers). Also, attention mechanism was utilized in three articles, either to improve recurrent networks [161] or in a transformer-inspired [172] solutions [138, 158]. Out of less commonly known approaches, multitask learning was utilized in four articles, either with fully connected (MLP) [62, 154, 160], convolutional [62, 154], or recurrent [154–156] networks. Zhang et al. [159] used a system consisting of a CNN network for mapping input to a low-dimensional feature space, followed by correlation-based feature extraction and Broad Learning System [173] for arousal and valence classification.

When creating a deep learning architecture, modifications can be applied in the form of person-specific layers (Tab. 4.1), which can be utilized for personalized data processing or reasoning. Yu et al. [138] first trained the general model and later added personalized layers parallel to the pre-trained (frozen) ones. By concatenating tensors obtained from the two streams, they created a representation containing both personal and general features and used personalized final layers for reasoning. Subject-specific layers can also be introduced at a later stage, only for personalized reasoning from data processed in a general way [62, 154–156].

Some articles researched interesting qualities and behaviors of deep learning algorithms, in addition to their accuracy in classifying affective states. Cheah et al. [140] conducted experiments on 1- and 2-path architecture to explore potential differences between them. Yu et al. [138] developed the modality fusion network, capable of stress classification using two physiological signals together or only one of them in case of a sensor malfunction. Li and Sano [155] researched whether it is better to fuse different modalities before or after feature extraction with an autoencoder. In another paper [156], they also researched different parameters affecting model quality, such as the number of clusters used for personalization, prediction quality in subject-

dependent and subject-independent scenarios, as well as the impact that utilized architecture has on autoencoder's reconstruction loss. Finally, Zhang et al. [159] conducted an ablation study and explored the impact of learning intra-modality features, using correlation-based features and Broad Learning System on classification results.

Although they were not used in any of the reviewed articles, other approaches to personalized architectures exist. Such methods are widely used, e.g., in natural language processing, and they may be a great source of inspiration for affective computing scientists working with time series. They are often implemented as simple modifications to deep learning architectures, such as trainable variables or embeddings modeling human bias (tendency to experience certain emotions) [174], or low-rank adapters (LoRA, [175]) learning personal adaptation functions [176].

4.1.4 Personalized tasks

In researched articles, authors most commonly studied stress and emotion recognition (Tab 4.3). Stress recognition was typically performed on its general definition, although Can et al. [162] focused on frustration, and Gasparini et al. [165] focused on cognitive load. We also include panic attacks [61] in this category, as we see them as abrupt and intense stressful episodes. Emotions, on the other hand, were divided into arousal, valence, and categorical ones. Out of articles utilizing arousal and valence, in three articles, authors decided to interpret the original labels in terms of emotion categories [76, 100, 141, 152]. Moreover, in three papers, emotion categories could be directly transformed into axes or quadrants of arousal-valence space [64, 99, 146]. Some other studies also researched mood [65, 154–156, 160] (all of them used data from the SNAPSHOT study [139]).

Affective states were obtained in reviewed articles using one of three types of annotations: self-assessments, stimuli labels, or external annotations (Tab 4.3). Out of them, self-assessments were used the most, either as a direct source of ground truth or to validate if stimuli induced expected emotions. Only in two cases researchers used annotations provided by external observers [157, 166].

Regardless of the origin, data also had to be translated into problems solvable using machine learning. Problems present in relevant articles can be divided into three main categories: binary classification, multiclass classification, and regression. The most commonly used (23 papers) was the binary classification problem. A unique (for this review) approach was presented by Rubin et al. [61], who treated binary

Table 4.2: Machine learning approaches in reviewed studies. Legend: *anomaly detection, ML - machine learning, SA - self-assessment, RF - random forest, SVM - support vector machine, MLP - multilayer perceptron, CNN - convolutional neural network, LSTM - long short-term memory, ? - deduced by us.

Category	Approach	Details	Used by
Machine learning problem	Binary classification	SA labels	[138, 141, 142, 144, 145, 148, 150, 158–162, 164][61]*[64]?
		Stimuli labels	[62, 147, 151–153, 162, 165]
		Stimuli labels verified with SA	[99, 100]
	Multiclass classification	SA labels	[76, 100, 140, 143, 148, 154, 158, 159, 162]
		Stimuli labels	[142, 145, 147, 149, 151–153, 165]
		Stimuli labels verified with SA	[99, 100, 146]
		External annotation	[166]
Regression	SA labels	[65, 154–156, 163]	
	External annotation	[157]	
Machine learning model	Classical	E.g., RF, SVM, 1-layer neural network	[61, 76, 99, 100, 141–146, 148–150, 152–154, 157–166]
	Deep	MLP	[62, 154, 160, 162]
		Convolutional models	[62, 64, 65, 140, 147, 151, 157, 159]
		Recurrent models	[100, 154, 157–159, 161]; [155, 156] - autoencoders [154] - CNN + LSTM [138, 158] - transformer-based models [159] - CNN + correlation-based features + Broad Learning System

Table 4.3: Affective states researched in reviewed studies.

Category	Affective state details	Used by
Emotions	Valence	[64, 76, 99, 100, 140, 141, 144, 145, 148, 151, 152, 157–159, 161]
	Arousal	[64, 140, 144, 145, 148, 151, 152, 157–159, 161]
	Categorical	[64, 99, 146, 153, 166]
Stress	-	[62, 65, 138, 143, 147, 149, 150, 154–156, 160, 163, 164]
	Frustration	[162]
	Cognitive load	[165]
	Panic	[61]
Other	Mood	[65, 154–156, 160]

classification (namely pre-panic vs. non-pre-panic states) as an anomaly detection problem. Next was the multiclass classification problem (to either three or four classes), which was considered in 20 papers. Regression of different levels of affective states was considered in five papers, out of which two performed it using MTL regression models [154–156].

4.2 Signals for personalized ER

Most of the reviewed articles conducted their research using physiological signals (Tab. 4.4). Out of these, heart activity and electrodermal activity measures were the most commonly used, with 15 and 14 cases of use, respectively. The least popular was affect recognition using speech and muscle electric activity (both used in only one case). Overall, signals used in personalized emotion recognition do not differ from those used in general emotion recognition (Chap. 2).

In 12 articles (35%) authors focused only on a single modality, choosing either cardiac activity [145, 146, 165], EDA [147, 150], EEG [64, 140, 148, 153, 161], or audio signals [166]. In articles where multiple modalities were explored, the most popular combinations included: (1) cardiac and electrodermal activity, together with body temperature [149, 158, 159, 164]; (2) cardiac and electrodermal activity [62, 138, 144, 151, 152, 157, 162, 163]; (3) electrodermal activity and

Table 4.4: Signals used for affect recognition in reviewed studies. Legend: * - also activity recognition from phone.

Signal category	Signal type	Used by
Physiological	Heart activity (among others ECG, PPG)	[61, 62, 76, 99, 100, 138, 142–146, 149, 151, 152, 157–159, 162–165]
	Electrodermal activity (EDA)	[62, 65, 138, 141, 144, 147, 149–152, 154–160, 162–164]
	Brain electric activity (EEG)	[64, 140, 141, 148, 153, 161, 163]
	Muscle electric activity (EMG)	[163]
	Respiration	[61, 163, 164]
	Temperature	[61, 65, 149, 154, 155, 158, 159, 164]
Behavioral	Accelerometer, gyroscope	[65, 99, 100, 142, 154–156, 162]
Audio	Speech	[166]
Other	Mobile phone data	[142, 143, 154, 160]; [76]*

body temperature [65, 154–156]; or (4) electrodermal activity and accelerometer signal [65, 154–156, 162].

We also noticed that utilized physiology and activity measures depended on a task of interest. In emotion recognition studies, authors often utilized heart activity together with electrodermal activity [144, 151, 152, 157–159], accelerometer data [99, 100], or by itself [76, 145, 146]. Also, brain electric activity was often used [64, 140, 141, 148, 153, 161]. Additionally, the only article that used speech features also focused on ER [166]. Although in stress recognition combining cardiac and electrodermal activity data was no less common than in emotion recognition [62, 138, 149, 162–164], studies that utilized EDA without heart activity were much more common [65, 147, 150, 154–156, 160], and only one article researched it using solely heart function features [165]. As mood-focused studies always researched it alongside stress [65, 154–156, 160], utilized signals were similar between the two tasks, although authors never utilized cardiac features, and in two out of five studies used mobile phone data [154, 160]. Note that this variation may be explained by the fact that all articles researching mood used the same SNAPSHOT study [139] in their

research, and none used cardiac features¹.

4.3 Validation in personalized ER

Assessing the quality of machine learning models for personalized affect recognition requires properly designed validation. Procedures and metrics should always be selected in regards to the solved problem and assumed approach. In personalized machine learning validation should be designed very carefully, as it is easy to cause data leaks from test to training set, resulting in incorrect (overestimated) results.

In reviewed papers we identified three validation components that were further divided into specific approaches, namely: (1) three categories of validation procedure, (2) six commonly used metrics and seven less popular quality measures, and (3) an information if authors compared performance of their personalized models with that of general ones. Out of all articles, one did not provide any clear explanation of used validation procedure [164].

4.3.1 Validation procedures

Most of the reviewed articles (26) utilized strategies where data was split over collected samples (Tab. 4.5). In five articles, authors utilized a simple holdout split, in 10 k-fold cross validation, in six a combination of both, with a holdout set used for final testing, and in two monte-carlo cross-validation. Also, some authors presented their own unique approaches, with Rubin et al. [61] who created holdout train-validation-test splits and swapped validation and test sets to perform experiments twice, and Wu et al. [64] who in one of their experiments divided the dataset into multiple non-overlapping sets on which algorithm was trained, and tested it on randomly selected subset of remaining data (we call it cross-training with random validation).

Although cross-validation typically produces more robust measures of the model's abilities to generalize than using just a holdout test, doing a simple split over data samples may result in knowledge leaks, regardless of the specific method used. In reviewed articles, examples of such incorrect methods include (to our best understanding) articles by (1) Tian et al. [161] who created clusters based on personality before splitting subjects into different datasets; (2) Quiroz et al.[99] using 10-fold cross-validation (random split) on already windowed data with overlapping windows, which probably caused a serious knowledge leak from the validation

¹Also, in dataset description provided by Li and Sano [156], no heart-related signal is mentioned

Table 4.5: Approaches to validation in reviewed studies. Legend: CV - cross validation, LkDO - leave k days out, LkSO - leave k subjects out, ARI - adjusted random index, PCC - pearson correlation coefficient, val - validation * - R understood as an average residual value in regression task, ★ - validation strategy used only for model optimization; ? - deduced by us.

Validation component	Approach	Used by
Validation procedure	Split over samples	[61, 62, 64, 140, 142, 147, 148, 151, 152, 157, 160] - holdout, train-(val)-test split [99, 100, 141, 143, 145, 146, 153, 159, 161, 162]; [61, 62, 65, 140, 148, 157, 160]★; [158]? - k-fold CV; [149, 166] - monte carlo CV; [64] - cross-training + random validation
	Split over subjects (people)	[65, 99, 142, 144, 146, 147, 150, 152, 155, 156, 161, 163, 165] - LkSO CV
	Split over time	[76, 138] - holdout [142, 150, 156]; [155]? - LkDO CV [144] - Leave-k-sessions-out CV [154]? - LkDO monte-carlo CV
	No info	[164]
Quality measures	Accuracy	[64, 76, 99, 100, 140–147, 149, 151–153, 159–162, 165, 166]
	F1-measure	[61, 99, 138, 142, 147, 151, 154, 158, 159, 165, 166]
	Precision	[61, 147, 151, 164, 166]
	Recall	[61, 147, 151, 158]; [164] - fixed recall
	AUC	[62, 99, 145, 160]
	MAE	[65, 154–156]
	Other	[61, 138] - Confusion matrices; [62] - Cohen’s Kappa?; [142] - ARI; [163] - 1-R*; [157] - RMSE, PCC, CCC
Comparison with general models	-	[64, 65, 99, 100, 141–144, 146, 147, 150, 152, 159, 162, 164–166]

sets; (3) Sah and Ghasemzadeh [147] who most probably normalized data before creating training and test sets, thus violating the assumption of training and test set independence; and (4) Gupta et al. [141] who performed model tuning using 10-fold cross-validation despite not having a separate test set (possibly resulting in a test set overfit). In total, 15 of reviewed articles utilize methods that could result in knowledge leak and provided descriptions do not allow rejecting such possibility [62, 64, 99, 100, 140, 141, 147, 149–151, 157, 158, 161–163].

Splitting datasets over subjects (people) is often used to avoid knowledge leaks. Although it is not flawless, it can be used and understood easily, especially when creating general models for the population (as done by 11 out of 17 articles comparing personalized and general models). Out of the articles that employed leave-k-subjects-out (LkSO) cross-validation, it may be utilized in a group-personalized setting based on similarities between people, as shown by, e.g., Hernandez et al. [150] who weighted model parameters based on subject similarities (although, they probably calculated weights incorrectly, leading to knowledge leak), or Tervoren et al. [142] who created reasoning models for groups of similar participants (based on questionnaires) and validated them by testing their ability to generalize to a new person belonging to the group.

Another method used to avoid knowledge leaks, used in five articles, was to perform data split over time, using either a holdout split [76, 138], leave-k-sessions-out cross-validation [144], or leave-k-days-out (LkDO) cross-validation [142, 150, 156], with Yu et al [154] probably using its monte-carlo version. If on-subject dependence is allowed (e.g., using personal models), validation based on time, especially when respecting data sequentiality, is a good way of measuring the model's prediction capabilities for new data. Because it controls for the day-wise specificity, this approach is suited only for data collected from repeated observations of people over a period of time (typically field studies), though when modified to shorter time periods, it may accommodate other factors, e.g., in a study with multiple stimuli presented in sequential order, or multiple sessions of experiment, validation over time of the study will be similar or even synonymous to leave-k-stimuli out (leave-k-sessions-out) validation procedure.

4.3.2 Metrics

The most commonly used metric was accuracy, utilized in 22 papers. Also, in some articles, it was the only used metric, despite a clear imbalance present in the

data [76, 152]. In three papers [147, 162, 166], although metrics other than accuracy were also used, they were provided only for some cases. F1-measure was used in 11 articles, precision and recall in five (in one [164] precision was evaluated at the fixed value of recall). Three papers utilized area under receiver operating characteristic curve (AUC), two papers used confusion matrices in addition to other metrics [61, 138], and Saeed et al. [62] as the only ones used Kappa coefficient (we deduce that it is the same as Cohen's Kappa).

Mean absolute error (MAE) was provided in three out of five articles trying to solve regression tasks. Xu et al. [163] provided regression accuracy understood as $1 - R$, with R being the average difference between ground truth and predicted value in regression task (possibly MAE). Other metrics used for regression included root mean square error, Pearson correlation coefficient, and concordance correlation coefficient [157].

4.4 Discussion

This review focused on reviewing methods for personalized affect recognition. During the study, we found three main areas in which the articles differed: (1) used strategies of personalization, (2) utilized input data (signals), and (3) employed strategies of validation. We further divided personalization strategies into subfields regarding (a) the subject of personalization, (b) personalization procedures, (c) algorithms and deep architectures used, and (d) task solved. Also, validation strategies differed between articles in terms of: (a) used procedures and (b) employed metrics.

4.4.1 Personalization subjects

Making a decision about a personalization subject requires some careful considerations. Its choice influences further decisions regarding utilized personalization procedures, models, and, possibly, validation. If research is focused on creating individualized solutions, training big models may be impossible, as, in general, it is hard to collect enough per-person data. Although some articles tried to utilize multitasking or transfer learning to alleviate low-data problems, utilizing such models in real life would prove hard, as they still require retraining entire models or some of their layers for each person. Although most group models are usually not concerned with those problems, they require knowledge and research on factors that impact emotions experienced by participants. In the reviewed articles, the authors utilized factors such as gender, answers to Big Five Personality Model questionnaires, Perceived Stress Scale, or even physiology to form groups of similar

people. In our opinion, further research is needed, as those methods do not seem to exhaust the scope of possible factors (e.g., those described in Sec. 3.4). Also, some articles created groups at random or based on the experimental group they belonged to. These methods may prove useful but are affected by the same problems as fully-personalized ones.

4.4.2 Training procedures and machine learning models

Among procedures used to create machine learning models, the most common was to train separate models for each personalization subject from scratch, either for a person or a group. Shi et al. [164] improved such models by adding personalized parameters to Other approaches also utilizing general knowledge about patterns present in the data were much less popular, with four articles using transfer learning, and personalizing some existing general models; one retraining previously created model; five using multitask learning to expand the amount of data used for training while designating task-specific heads to predict personal or group affect. The most unique was the approach described by Gasparini et al. [165], which focused on processing personal data to eliminate personal specificity from physiology and create general models using such data.

Despite much progress in deep learning, most of the research utilized classical machine learning algorithms, such as SVM, Random Forest, or multilayer perceptrons. It may be caused by the fact that most recent deep learning algorithms require vast amounts of training samples to achieve good accuracy, and in affective computing, the number of per-person samples is usually low. It would also explain why some of the works trained deep learning algorithms using transfer learning (learning using some previous knowledge as a foundation) or multitask learning procedures (using data from all subjects to train a general part of a model while also fitting personalized layers or parameters). Out of articles that used deep learning models, algorithms based on convolutional and recurrent layers were similarly popular. Also, two more recent articles achieved promising results using transformer-like models, showing that although they, in general, require large amounts of data, personalization of deep learning models is a promising direction.

4.4.3 Affective tasks

Most of the reviewed articles focused on emotions, either as levels of arousal or valence or as discrete named states, such as anger or joy. Other research focused on stress and mood. In most cases, participants' inner feelings annotated using

questionnaires were used as a source of ground truth, followed by some articles using stimuli categories, and two studies [157, 166] employing external experts to annotate emotions of experiment subjects. Before using them as output variables, these affective states had to be transformed into categories (classification) or levels (regression). Most commonly, researchers used binary categories, and multiclass classification was slightly less popular. Regression was considered in only six papers.

The popularity of binary classification may suggest, especially if we consider the general unwillingness to publish negative results [177–179], that affective state recognition is a difficult task, regardless of its nature (short- or long-lasting states). Additionally, we observed that the proportion of studies utilizing self-reported emotion, as opposed to stimuli labels, is noticeably higher for binary classification than for the multiclass problems (15 to 9 vs. 9 to 11). This may suggest, even more so if we consider all of the subjectivity that people’s emotions and mental processes exhibit, that recognition of self-rated emotions is more difficult than recognition of emotions targeted in an experiment. For example, if a stimulus was prepared to induce fear or anger, it may cause elevated arousal, which some participants may regard as the targeted emotion, while others may associate with, e.g., excitement. The low popularity of regression adds to this conclusion, as regression of precise emotion values requires more precise links between annotated values and input signals. If those links differ between conditions because of momentary subjective perception, it may be difficult or impossible to precisely model levels of emotion. However, these observations further add to the importance of person-focused processing, as opposed to the generalized one.

4.4.4 Utilized signals

In reviewed articles, the most popular signals were those of physiological origin. Out of them, the most common among different cases were signals coming from heart activity and electrodermal activity, with 16 and 15 cases of use, respectively. The least popular was recognition using speech and muscle electric activity, both used in only one case. Also, the signals differed between the different tasks that the researchers focused on. In emotion recognition studies, popular signals or their combinations included: (1) heart activity together with electrodermal activity (6 articles), accelerometer data (2 articles), or by itself (3 articles), and (2) brain electric activity (6 articles). Moreover, speech features were only used for ER [166]. In stress recognition, combining cardiac and electrodermal activity data was as

popular as in emotion recognition (6 articles), but studies that utilized EDA without heart activity were much more common (7 articles), and only one article used only features from cardiac signal [165]. Mood was always researched alongside stress (5 articles), but authors never utilized cardiac features which were common in stress research. This difference is probably caused by characteristics of SNAPSHOT study [139], which was a source of data for all of those articles.

4.4.5 Validation for personalized emotion recognition

In general, the assumed training procedure and the amount of subject dependence are two main factors restricting the possible scope of validation procedures. The validation procedure should control all factors that may impact the model results, but specifics can differ between experimental designs. If research focuses on creating general or group-specific (e.g., personality- or gender-aware ones) models, testing them on an unseen set of subjects, or unseen stimuli for laboratory data, may be a preferred approach. In reviewed articles, the popular leave-k-subjects out (LkSO, often $k=1$) validation strategy was used for training general models in 11 out of 17 of them, and for training personalized (group) models in 11 out of 19 articles (including the hybrid approach to subject personalization). Out of articles considering time-based validation, one used splits based on sessions in laboratory [144], two utilized holdout test sets [76, 138], and others used leave-k-days-out (LkDO) procedure (all of them were studies conducted outside the laboratory).

Most reviewed studies used simple splits based on data samples, with the holdout, k-fold, or monte-carlo validation procedures. Such procedures may be correct if splitting is done on independent samples, but if performed on dependent samples, high accuracies of obtained models may be caused by too-far-reaching dependencies in collected data rather than the validity of the personalization procedure. An extreme example of such an approach would be dividing a short annotated recording (of a few seconds or minutes) into overlapping windows and randomly dividing it into training and test sets, resulting (most probably) in a non-independent test set and yielding misleading results (as Quiroz et al.[99] probably did). Of course, samples collected from one person will always exhibit some dependencies, but if the samples themselves were not collected during one short-lasting trial, one could assume that all dependencies are coming from general tendencies and patterns in this person's signals rather than from a particular momentary state. Validation misconduct leading to overestimated generalizability of methods may also take other forms, e.g., adjusting models using test set resulting in overfitting [141], creating

clusters of similar people before splitting data into training and test sets [161], or normalizing data using statistics from the whole dataset [147]². Therefore, each study requires careful consideration of the validation procedure, which should be precisely stated and explained together with a rationale. Unfortunately, most authors ignored this aspect in their articles, either providing very high-level descriptions or not providing any details at all [164].

The use of classification metrics in reviewed studies with accuracy often used as the only metric is worrisome, as it shows that many authors did not consider the effect that imbalanced datasets may have on trained algorithms and how obtained results may be misleading. When combined with a binary classification problem and a simple holdout or k-fold split, in the event of a high imbalance between classes, it is easy for the algorithm to fit the majority class and achieve high accuracy, leading to overestimated model capabilities. This further adds to the conclusion that the importance of proper validation is often underestimated. An example of well-considered imbalance are articles by Yu et al. [138, 154], where they not only used the focal loss to account for imbalance during training but also reported F1 measure as well.

²Note that all of those examples are based on our understanding of reviewed articles, and our understanding may be affected by insufficient descriptions of validation procedures.

Chapter 5

IN-THE-LABORATORY DATA COLLECTION

To perform analyses and train machine learning models, one needs data - generally, the more, the better. Collecting this data poses many difficulties, especially when its quality depends on answers provided by people. The complicated process of obtaining data may become really arduous when employed people are not fully capable of providing certain and reliable answers or annotations. Emotion research is one of the fields that are most affected by such factors, as emotions are not only subjective but can also be very confusing and hard to accurately estimate, not to mention annotating them in a precise and repetitive manner.

There are many issues with existing datasets for emotion research, the main one being the lack of open datasets collected in everyday life. Laboratory datasets still dominate the field because of their signal quality, known study conditions, and the cost of such study. However, recently, portable devices suitable for everyday life have improved to the point where collected physiological measures are accurate enough to be useful. Because of that, researchers started conducting emotion research in real life, but the number of available datasets is still very limited. With all of that in mind, we (Emognition) decided to collect our own psychophysiology datasets, first in the laboratory and later in real life.

Aware of our limited experience, we decided to first carry out a laboratory study with devices suitable for real-life use. This way, we could face at least some of the difficulties occurring when collecting affective data, learn from them, and better prepare for in-the-wild experiments that we later conducted. This chapter focuses on collecting data from people and on the issues that arise during the process.

Contents of this chapter originate from the co-authored article, published in a peer-reviewed journal:

- [24] S. Saganowski, J. Komoszyńska, M. Behnke, B. Perz, D. Kunc, B. Klich, Ł. D. Kaczmarek, and P. Kazienko, “Emognition dataset: Emotion recognition with self-reports, facial expressions, and physiology using wearables,” *Scientific data*, vol. 9, no. 1, p. 158, 2022. DOI: 10.1038/s41597-022-01262-0.

In the original article [24], I contributed mainly to data curation, its investigation, and technical validation. Additionally, I personally wrote the lessons learned section (Sec. 5.3.2) solely for this dissertation.

5.1 Experimental procedure

Once seated in the laboratory, participants were instructed about their role in the experiment and asked to abstain from performing excessive movements (e.g., swinging on the chair) and to ensure that their faces were visible to the camera. We also informed them that if they felt uncomfortable during the experiment, at any moment, they could skip any stimulus or quit the experiment. For the experiment, participants were not directly watched by the researcher. However, they could ask for help at any moment.

First, participants watched a black screen with dots and lines displayed (washout) for 5 minutes. During that time, we collected their physiological baseline and self-reported emotions after the clip ended (emotional baseline). For the actual experiment, participants were exposed to 10 videos inducing emotions of amusement, anger, awe, disgust, enthusiasm, fear, liking, surprise, and sadness, and one video targeting a neutral state, with durations ranging from 49s to 121s. Videos were organized in the randomized circular queue order, and for each of the participants, they were displayed starting at one item later in the queue than for the previous participant. Also, each stimulation was preceded by the washout video lasting two minutes.

After watching a stimulus, they were tasked to provide ratings of their emotional experience using two types of self-assessments: ratings of discrete emotion items and levels of emotion components. Additionally, at the end of the experiment, we asked participants to fill out an additional questionnaire and provide us with (1) details about video clips they were familiar with before the study and (2) any remarks they had about the experiment. After completing the whole procedure, each participant received a 50 PLN voucher for the online store.

5.1.1 Physiological data

During all inductions, multiple measures of subjects' PNS activity were collected using three unobtrusive devices: Empatica E4, Samsung Galaxy Watch, and Muse 2 headband. Additionally, video of the participant was recorded using a Samsung Galaxy S20+ smartphone. Details on all collected measures are presented in Tab. 5.1.

Table 5.1: Measures available in the Emognition dataset (from [24]).

Collection method / device	Available data
Emotion assessments	<ul style="list-style-type: none"> • Ratings for dimensional model of emotions: arousal, valence, and motivation • Ratings for discrete model of emotions
Empatica E4	<ul style="list-style-type: none"> • Blood volume pulse (BVP) • Interbeat interval (IBI) • Electrodermal activity (EDA) • 3-axis accelerometer • Skin temperature
Samsung Galaxy Watch	<ul style="list-style-type: none"> • Heart rate (HR) • Peak-to-peak interval (PPI) • Raw blood volume pulse (BVP) • Processed blood volume pulse (BVP) • 3-axis accelerometer • 3-axis gyroscope • 4-axis rotation
Muse 2	<ul style="list-style-type: none"> • Data from AF7, AF8, AF9, and AF10 electrodes: <ul style="list-style-type: none"> – Raw EEG signal – Absolute powers for Alpha, Beta, Gamma, Delta, Theta frequency bands • 3-axis accelerometer • 3-axis gyroscope
Samsung Galaxy S20+ 5G	<ul style="list-style-type: none"> • Upper-body video recording

We used a custom application running on a computer to collect data on emotions. This app utilized Empatica E4 SDK, so data from the Empatica E4 wristband was synchronized with the stimuli out-of-the-box. On the other hand, data from Samsung Watch and Muse 2 devices were collected separately via smartphone using third-party apps, and they had to be synchronized with the participant's session. Because each device has its own CPU time, we could not use signal timestamps to synchronize them. Instead, we utilized a simple protocol: once turned on, all devices were put on a table, which was hit with a fist after a few seconds. It resulted in a visible peak in ACC signals from all devices, which was used to match their time to Empatica E4 time.

5.2 Results

Our laboratory dataset was collected between July 16 and August 4, 2020, at the Wrocław University of Science and Technology. It consists of data from 43 subjects (21 females, M age = 22.37 years, SD = 2.25). For each participant, we released data from all experimental stages (stimulus presentation, washout, self-assessment) and all devices (Muse 2, Empatica E4, Samsung Watch) on Harvard Dataverse Repository [180]. The released dataset consists of:

- recordings from 3 devices when watching 10 film clips, 3 phases each (washout, stimulus, self-assessment);
- baseline recordings from between video clips (3 devices, 2 phases: baseline and self-assessment);
- self-assessment responses, the control questionnaire, and metadata (e.g., demographics and information about wearing glasses)

The types of available physiological and video data are illustrated in Fig. 5.1a. Additionally, for each video frame, we provide annotations of facial expressions obtained using OpenFace (facial landmark points and values of action units, Fig. 5.1c) and Quantum Sense (values of six basic emotions and head position) packages. Other information regarding the dataset structure (e.g., file naming conventions and variables available in each file) can be found in the dataset’s README.txt file, which is available in the repository.

5.2.1 Technical validation

To examine the effect that film clips had on participants (elicited emotions, within conditions test), we used repeated-measures analysis of variance (rmANOVA) with Greenhouse-Geisser correction and calculated recommended effect sizes of η_p^2 for the tests [181, 182]. To check differences in experienced emotions between conditions (e.g., if the level of self-reported anger in response to the anger stimulus was higher than in response to the other stimuli), we calculated pairwise comparisons with Bonferroni correction of p -values for multiple comparisons.

As summarized in Fig. 5.2 (also Appendix Tab. B.1 and Tab. B.2), watching film clips evoked the targeted emotions, and differences in self-reported emotions in film clips can be interpreted as large [183]. Pairwise comparisons indicate that self-reported targeted emotions achieved the highest values for stimuli targeting

Table 5.2: Signal-to-noise ratios (SNRs) statistics. Columns inherit their names after abbreviations of physiology measures (EEG – electroencephalography, BVP – blood volume pulse; TEMP – skin temperature; EDA – electrodermal activity). All values are in decibels (dB), except for count. Q_x are values for quantiles, where x denotes the percentage of samples falling into the bin from min to Q_x, e.g., Q50 is the median. Differences in samples used for calculations (count) come from the malfunctioning of the devices (from [24]).

	MUSE (EEG)				Samsung	Empatica		
	TP9	AF7	AF8	TP10	BVP	BVP	TEMP	EDA
count	1312	1312	1312	1312	1344	1350	1348	1350
mean	36.22	37.74	37.12	37.16	33.77	33.43	26.81	26.66
std	6.02	9.44	11.13	6.77	2.86	2.27	3.04	2.81
min	0.88	5.85	4.13	4.84	17.76	23.57	3.01	18.50
Q0.3	5.15	8.74	7.26	7.39	26.24	25.51	17.48	19.78
Q25	33.74	37.50	38.18	35.31	31.40	32.12	24.46	24.36
Q50	37.46	40.84	41.39	39.12	35.46	33.54	28.56	27.90
Q75	40.17	43.47	43.99	41.12	35.82	34.93	28.84	28.68
max	45.63	48.62	49.88	46.73	42.91	40.83	36.91	37.00

them (e.g., self-reported anger in response to the angry film clip). Furthermore, we observed the high intensity of some emotions in more than one film clip. Similarly, some film clips elicited more than one (targeted) emotion. Such effects are frequently reported for emotion elicitation procedures [184–186].

To validate the quality of collected physiological signals, we used signal-to-noise ratios (SNRs) obtained by computing signal autocorrelation and fitting the second-order polynomial to the resulting values. It was done separately for all physiological recordings (all participants, baselines, film clips, and experimental stages). We excluded accelerometer and gyroscope signals from SNR calculations, as all experiments were conducted in a sitting position. Also, as we discarded samples corrupted due to sensor malfunction, sample counts in Tab. 5.2 differ between signals.

SNR statistics indicate that signals are of high quality (Tab. 5.2). Mean SNR ranged from 26.66 dB to 37.74 dB, with standard deviations between 2.27 dB and 11.13 dB. The minimum observed SNR was 0.88 dB for the ECG signal from the TP9 electrode, but only 0.3% of signals from this electrode had SNR values below 5.15 dB. Moreover, for other EEG electrodes, the computed SNR was never below 4.13 dB.

Additionally, we analyzed Quantum Sense annotations (six basic emotions; neutral, anger, disgust, happiness, sadness, surprise) to check the software’s emotion recognition quality. In general, it performed well within conditions but poorly between

conditions. Wrong or missing annotations occurred most often when participants covered their faces with a palm or leaned toward the camera. Also, we noticed that participants' faces usually remained neutral for the majority of the experiment, and they expressed emotion very briefly during intense moments in the stimuli. In some cases, participants familiar with the stimulus (e.g., watched the source movie) reacted differently (e.g., smiled instead of expressing disgust). For a more detailed analysis, see Saganowski et al. [24].

5.2.2 Issues in collected data

During the data collection effort, we can expect unexpected situations to occur. They may originate from insufficient testing of research procedures, people not following the required protocol, or from unknown causes with a seemingly random nature.

Although all procedures had been tested beforehand, during the study, some technical problems still occurred with collecting data. In two cases, we observed undesirable behavior of the Empatica E4 wristband: on one occasion, it could not connect to the computer, and on another, it connected instantaneously (which was also unexpected). In both cases, restarting the device fixed the problem. Other issues with devices measuring physiology included the Samsung Galaxy Watch running out of battery (one participant), missing EEG data from Muse 2 (one participant), and a failed connection between computer and server, resulting in a substantial delay between stimulus and questionnaire (~8min 30s, one participant). Also, in the case of five participants, facial expression recordings were either corrupted or missing (possibly due to insufficient space on the smartphone).

Issues of non-technical origin included participants who failed to maintain the required posture during a study by leaning on the chair or covering their face (two participants), people who were already familiar with film clips used during the study (12 participants), or a person who came visibly intoxicated. Also, one participant reported trouble with immersing in shown film clips without knowing the full movie context, and three skipped some parts of inductions.

5.3 Discussion

The Emognition dataset described above can be seen as a step toward in-the-field research on emotions. Although the study was not conducted outside the laboratory, nor was its setup naturalistic, all of the devices used were unobtrusive and could be easily used in a real-life study. While all of them were designed with everyday use in mind, we note that the Muse 2 headband should be used in stationary conditions,

as the EEG signal is susceptible to artifacts from movement and nearby electric devices. Nevertheless, all of the devices are much easier to set up than their laboratory counterparts.

5.3.1 Data quality

When compared with other datasets for emotion research, Emognition with its 43 participants does not fall below the average population count for [16, 187]. Moreover, it exceeds the average number of elicited emotions per participant [187]. Also, we obtained approval of ethics committee before conducting the study and collected written consent from participants, which are not always considered in other studies [187].

The study can be seen as overall successful, according to the results of technical validation. Most of the signals exhibit satisfactory signal-to-noise ratios, with a minority of recordings falling below the often assumed acceptable level of 25dB (Tab. 5.2). Also, emotion inductions can be considered successful, with all film clips inducing the highest levels of targeted emotion between all film clips (Fig. 5.2a), and with targeted emotions being always strongest or near strongest within corresponding film clips (Fig. 5.2b).

Some of the film clips provide interesting examples for analysis. For example, in the video targeting anger, anger achieved the strongest ratings among all videos, but at the same time, disgust, sadness, and surprise were, on average, stronger than anger within this film clip. Also, awe was often induced not only during induction targeting it but also for enthusiasm and liking stimuli. It is also worth noting that registered levels of emotions are often below the middle value of the scale. Possible causes include (1) lack of immersion (e.g., due to insufficient context), (2) weak effect of video stimuli in general, and (3) referencing induced emotions to a very intense past experience. However, this effect is not present in self-reported values for valence, arousal, and dominance. Thus, determining the exact effect of emotion inductions requires further investigation, possibly including the specificity of particular subjects.

5.3.2 Lessons learned

Another takeaway from this work are lessons that we learned facing issues observed during this endeavor. We believe that they may be of value for researchers conducting studies on emotion psychophysiology in both laboratory and outside-laboratory settings.

First, only basic information about participants was collected during the study. While subjects may find short procedure pleasing, having detailed data on their demography, personality, health issues, or some other factors that influence emotion or physiological signals (e.g., medication used) could allow more detailed analyses and better understanding of collected dataset. Also, ECG signal was not collected during the study, despite its popularity among datasets for research on emotions [187] and availability of simple off-the-shelf ECG chest straps, such as Polar H10.

Problems with missing connection between devices, or between computer and the server collecting data could be addressed in various ways. First of all, connection should be thoroughly tested in terms of their reliability, especially during long recording periods. Also, operability of an application for conducting the study should not depend on internet connection, as it is crucial for obtaining reliable results. However, if it is impossible to design an app working entirely without internet connection, it should be limited to periods between different inductions, not between induction and self-report questionnaire.

Some of issues, such as insufficient charge of a smartwatch or lack of memory on the phone could be avoided if protocols were better designed. Such protocol should include detailed steps to follow before and after each participant's session. Ideally, for each device there should be backup available, in case of sudden malfunction.

Lastly, we learned a lot regarding the management of study participants. In some cases people showed up late, asked to change the date of visit right before their allocated time slot, or canceled their participation at the last moment. Moreover, some of participants showed visibly sick or intoxicated. We did not expect such situations, as we asked them to show up in their best health for the experiment (or reschedule if they required it). Appropriate well-considered protocol could also help in such situations, as it would specify steps to follow in such unusual situations.

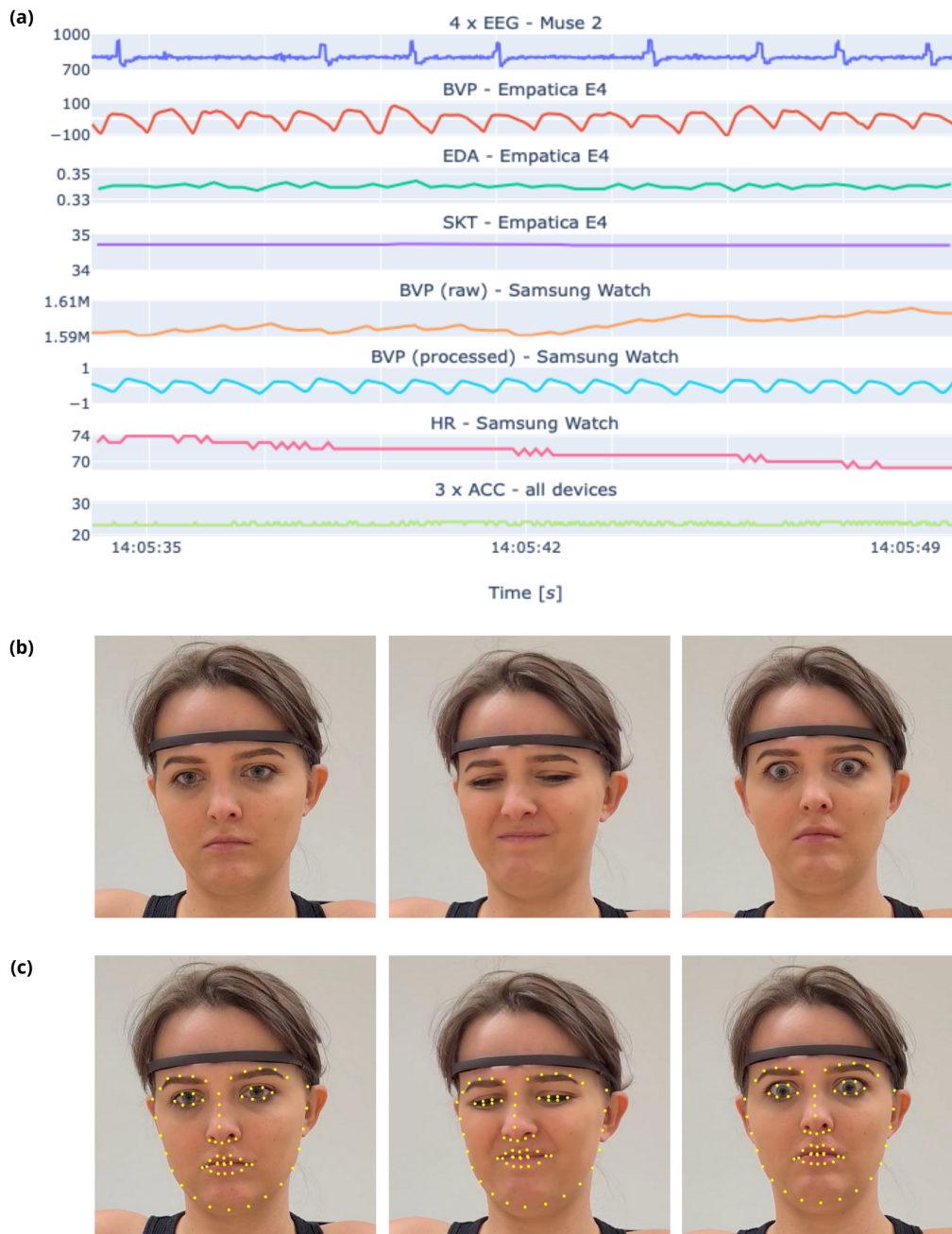


Figure 5.1: Examples of data available in the Emognition dataset: (a) physiological signals recorded with wearable devices: 4 x EEG (Muse 2); BVP, EDA, SKT (Empatica E4); raw BVP, processed BVP, HR (Samsung Watch); ACC from all devices; (b) upper-body recordings capturing the facial reactions to the stimuli, from the left: neutral, disgust, surprise; (c) facial landmarks generated with the *OpenFace* library facilitating emotion recognition from the face. The participant gave written consent to include her image in this article (from [24]).

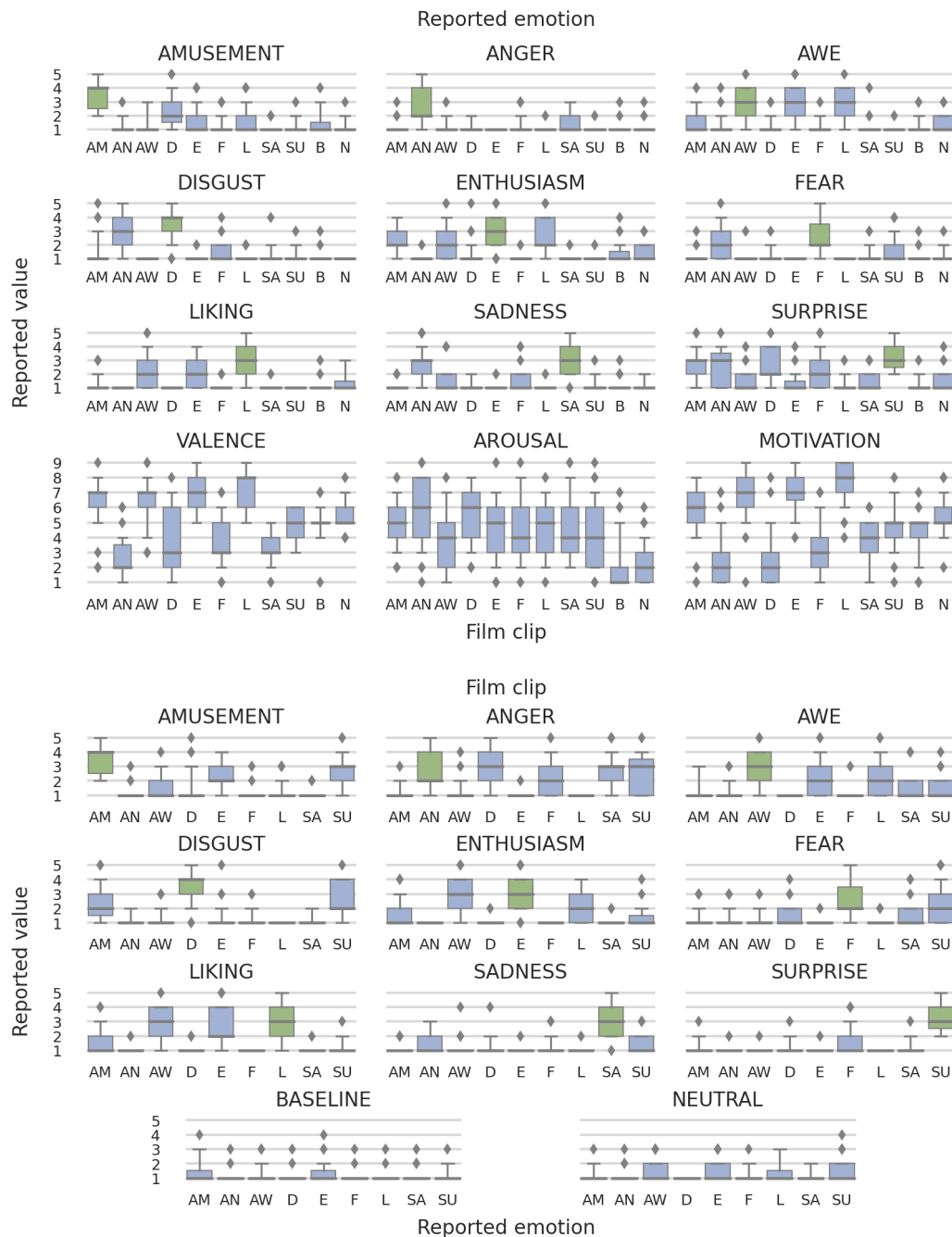


Figure 5.2: Distribution of self-reported emotions between conditions (upper), i.e., levels of emotions elicited by different films, and within conditions (lower), i.e., emotions evoked by different film types. Vertical axes denote emotion intensity (1-5 for discrete emotions, 1-9 for SAM). Horizontal labels represent film clips (upper), or discrete emotions reported by the participants (lower). Green color indicates targeted emotion. Boxes depict quartiles of distributions and whiskers the span from the 5th to 95th percentile (diamonds - outliers). Legend: AM - amusement, AN - anger, AW - awe, D - disgust, E - enthusiasm, F - fear, L - liking, SA - sadness, SU - surprise, B - baseline, N - neutral. From Saganowski et al. [24].

Chapter 6

COLD START AND GROUP PERSONALIZATION

Gathering annotated emotional events in everyday life is costly and challenging. In this section, we describe research, where we focused on the problem of reducing this complexity by building and employing machine learning models for detecting intense emotional states. We also consider the cold start problem, where researchers have no signals collected from the target subjects (users) at the beginning of a data collection experiment. We investigated the possibility of using per-group personalization to address this problem and the amount of data needed to perform this procedure.

Contents of this section originate from the co-authored article, published in peer-reviewed conference materials:

- [28] S. Saganowski, D. Kunc, B. Perz, J. Komoszyńska, M. Behnke, and P. Kazienko, “The cold start problem and per-group personalization in real-life emotion recognition with wearables,” in *2022 IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, WristSense 2022 - The Eighth Workshop on Sensing Systems and Applications Using Wrist Worn Smart Devices, Best Paper Award, IEEE, 2022, pp. 812–817. DOI: 10.1109/PerComWorkshops53856.2022.9767233.

In the original article [28], our contributions included designing the experiment, collecting data, and investigating the results, with special contributions of myself to creating and validating deep learning models for emotion detection.

6.1 Materials and methods

The data and methods presented below bear a resemblance to the LarField study, also described in Chap. 7. That is because this research was conducted when we were exploring different approaches and methods for conducting a study in real life. Results and conclusions described in this section were later used to improve the protocols, applications, and models used in our large field study.

6.1.1 Dataset

Data for experiments in this section came from two daily life studies (Study A and Study B) performed by the Emognition research group. The studies were similar,

Table 6.1: Distribution of data used in the research (from [28]).

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

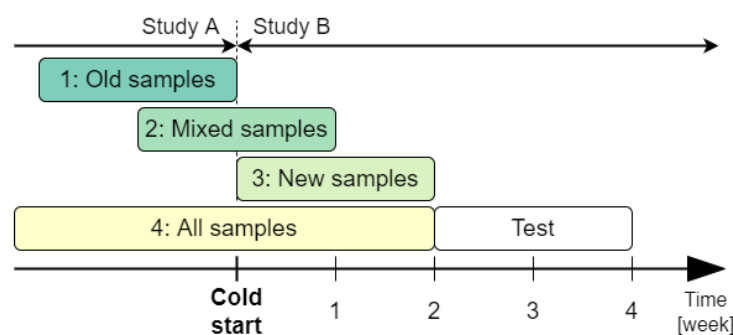


Figure 6.1: Scenarios considered in cold start and group personalization experiments (from [188]).

with the second having a slightly modified self-assessment and an entirely new cohort of subjects. Study A lasted about seven months and involved 11 participants (four females). Study B lasted two months and involved 13 participants (six females).

Study A focused mainly on collecting physiological signals during events in participants' everyday lives when they experienced intense emotions. The collected data were used to create an ML model for recognizing intense emotions in real-time [189]. Such a model was used for more efficient data gathering in part of Study A and in Study B.

Study B focused mainly on validating several models for predicting intense emotions, and it involved further data collection. Also, the newer version of the Emognition system used in Study B included a shorter self-assessment questionnaire and three types of assessment triggers. As the analyses were conducted during Study B, they consider only the first four weeks of Study B. Also, although in this study we had

13 participants (6 female), we used data from only the five most active participants (two females) who had the highest number of filled-out self-assessments.

The emotions were collected using the Emognition system [15], using short self-assessment questionnaires triggered at quasi-random times, at the request of the machine learning model recognizing intense emotions, or on demand. Same as in our LarField study (Chap. 7), participants were asked: (1) if they felt intense emotions (used to categorize emotions as intense emotions or neutral states), and (2) how they felt in terms of valence and arousal from 1 (extremely negative or sluggish) to 100 (extremely positive or aroused). Participants could also provide a free text with comments in the last field of the questionnaire.

In total, the dataset used in this experiment contained 1075 self-reports (440 intense emotions and 635 neutral states) collected in both studies (Tab. 6.1). We analyzed data from 16 participants (6 female) between 18 and 54 years of age ($M=26.86$, $SD = 8.29$). Participants 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.

6.1.2 Signal processing and feature extraction

For the experiments, we used six types of signals:

1. BVP signal from the smartwatch (raw)
2. BVP processed with median and band-pass Butterworth filters
3. Heart rate (from smartwatch)
4. PP-interval (from smartwatch)
5. Heart rate computed from PP-interval
6. Accelerometer data (ACC)

Depending on the model setup, we utilized different signals: all of the above signals (1-6) or all cardiac signals (1-5) for feature-based models, and cardiac signals apart from HR from PP-interval with (1-4 and 6) or without (1-6) accelerometer data for end-to-end (e2e) models. From collected signals, we extracted 140s-long windows centered around the annotated emotional event (the time when self-assessment

Table 6.2: Features extracted from physiological signals (from [28]).

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)

began or timestamp marked by the reasoning model embedded in the smartphone application). We discarded windows containing less than 90% of expected samples (as compared with sampling frequency multiplied by the length of extracted window). Next, all signals were resampled using *resample* function from SciPy[190] and further cropped to windows of 60s centered around the emotional event for each signal. These windows were further divided into three non-overlapping 20s-long parts to explicitly provide models with physiology before, during, and after the event and allow them to learn short-time psychophysiological dependencies.

Signal windows were either provided to models as is or were used for computing descriptive features (see Tab. 6.2). We used statistical features such as min, max, and mean values of the signal, or the signal's variance and standard deviation for each part of a window, and their differences between consecutive parts of a window (e.g., difference between minimum values in the first and second part of a window). We also computed features in the frequency domain, such as minimum, maximum, or average values in the power spectrum. Additionally, we computed mean value in low- and high-frequency power spectra of a BVP signal, and two date-related features. All features were concatenated (746 with ACC or 418 without it), and resulting vectors were used to train machine learning classifiers.

6.1.3 Models

Experiments were conducted using two categories of machine learning models: (1) feature-based models, i.e., AdaBoost, k-Nearest Neighbours (KNN), Random Forest, Support Vector Machine (SVM), Multilayer Perceptron (MLP), and a Fully Convolutional Network (FCN) with a unit kernel, and (2) end-to-end models, i.e., FCN, FCN with LSTM layers (FCN-LSTM), and Residual Neural Network (ResNet). FCN-LSTM architecture consisted of several FCN channels that supplied processed information to LSTM layers. In FCN and ResNet architectures, each part of a window for each input signal was treated as separate channels (e.g., 3 parts * 4 signals = 12 channels). The deep learning architectures were adapted from the article by Dzieżyc et al. [121] and programmed in PyTorch [191], whereas classical machine learning algorithms were implemented using scikit-learn [192].

6.1.4 Training procedure

When preparing scenarios 1, 2, and 4 (abbreviated as S1, S2, and S4, respectively), we balanced data by repeatedly sampling it at random five times. Each split was further divided into training and validation sets and used to tune hyperparameters. Scenario 3 (S3) did not require balancing, but we used monte-carlo validation repeated five times to obtain results independent of a particular training and validation split.

The hyperparameters were optimized separately for each scenario and model. For scikit-learn models, we used the grid search method, and for deep learning models, we used random search, as it is more efficient for long training [193]. In all cases, we chose the best set of hyperparameters based on the achieved F1-macro score and retrained such models on the whole data splits (merged training and validation data). Finally, models were tested on the previously unseen data from the last two weeks of Study B (see Fig. 6.1).

6.2 Results

The results of each scenario and model are presented in Tabs 6.3 and 6.4. We do not report standard deviations in S3, as we did not perform random subsampling in this scenario. The three considered metrics are: (1) F1 measure on class 1, as it measures models' capability to catch emotional events, which in our case are more important than neutral states; (2) macro averaged F1, as it measures the overall performance of the model (emotions and neutral states); and (3) accuracy as it is the most widely used quality metric.

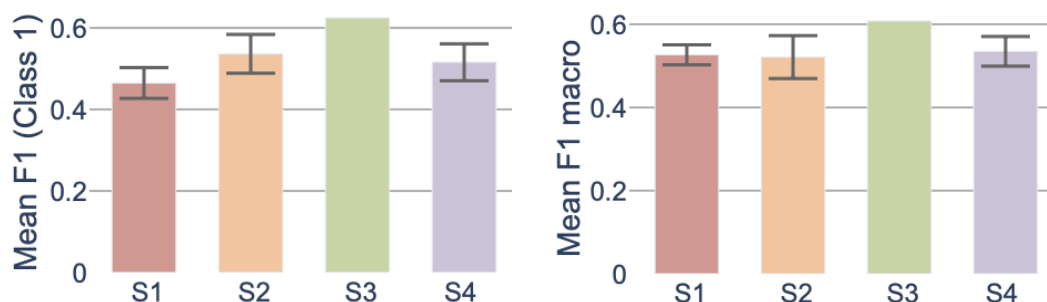


Figure 6.2: Mean F1 (Class 1) and mean F1 (Macro) scores for AdaBoost classifier for all scenarios (from [28]).

Overall, models trained in scenario 3 achieved better recognition quality than other scenarios. 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 (Fig. 6.2). AdaBoost and SVM models achieved the best results among classical feature-based approaches, MLP with ACC was the best feature-based deep learning model, and FCN-LSTM with ACC beat other e2e deep learning models. The mean differences between S1 (training only on general data) and S3 (training only on new data) are 0.09 in F1 on class 1, 0.05 in F1 macro, and 0.05 in accuracy, all in favor of S3.

These findings suggest 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 use of Friedman’s statistical test [194] to examine our results. It confirmed that models created in scenario 3 were the best, with S4, S2, and S1 following in the given order ($p = 3E-6$). Shaffer’s post-hoc [195] indicated statistically significant differences between the results of S1 and all other scenarios. However, it also indicated that differences between the results from other scenarios are statistically insignificant, i.e., S2 vs. S3, S2 vs. S4, and S3 vs. S4. Also, there was no clear improvement in models using accelerometer data compared with models without ACC.

Table 6.3: Results for investigated scenarios - classifiers using only cardiac data as input (metric \pm std). The highest scores for each classifier/architecture and performance measure are in bold (from [28]).

Model	Metric	S1	S2	S3	S4
AdaBoost	F1 class 1	0.47 \pm 0.04	0.54 \pm 0.05	0.62	0.52 \pm 0.05
	F1 macro	0.53 \pm 0.02	0.52 \pm 0.05	0.61	0.53 \pm 0.04
	Accuracy	0.54 \pm 0.03	0.52 \pm 0.05	0.61	0.54 \pm 0.04
KNN	F1 class 1	0.45 \pm 0.04	0.52 \pm 0.04	0.53	0.49 \pm 0.03
	F1 macro	0.53 \pm 0.03	0.53 \pm 0.03	0.53	0.55\pm0.01
	Accuracy	0.54 \pm 0.03	0.53 \pm 0.03	0.53	0.56\pm0.01
Random Forest	F1 class 1	0.48 \pm 0.05	0.60\pm0.03	0.58	0.60\pm0.03
	F1 macro	0.51 \pm 0.03	0.56 \pm 0.04	0.54	0.60\pm0.03
	Accuracy	0.52 \pm 0.03	0.56 \pm 0.03	0.54	0.60\pm0.03
SVM	F1 class 1	0.49 \pm 0.04	0.56 \pm 0.05	0.62	0.54 \pm 0.03
	F1 macro	0.53 \pm 0.03	0.52 \pm 0.03	0.57	0.56 \pm 0.04
	Accuracy	0.53 \pm 0.02	0.52 \pm 0.03	0.57	0.56 \pm 0.04
MLP	F1 class 1	0.51 \pm 0.06	0.54 \pm 0.05	0.57	0.48 \pm 0.03
	F1 macro	0.53 \pm 0.05	0.55 \pm 0.06	0.57	0.50 \pm 0.03
	Accuracy	0.53 \pm 0.05	0.55 \pm 0.06	0.57	0.50 \pm 0.03
Resnet e2e	F1 class 1	0.44 \pm 0.11	0.50 \pm 0.15	0.58	0.55 \pm 0.02
	F1 macro	0.51 \pm 0.03	0.54 \pm 0.05	0.62	0.55 \pm 0.03
	Accuracy	0.53 \pm 0.02	0.57 \pm 0.02	0.63	0.55 \pm 0.03
FCN e2e	F1 class 1	0.51 \pm 0.05	0.53 \pm 0.13	0.63	0.56 \pm 0.02
	F1 macro	0.54 \pm 0.04	0.52 \pm 0.06	0.64	0.58 \pm 0.01
	Accuracy	0.54 \pm 0.04	0.54 \pm 0.04	0.64	0.58 \pm 0.01
FCN-LSTM e2e	F1 class 1	0.45 \pm 0.05	0.55 \pm 0.05	0.56	0.61\pm0.02
	F1 macro	0.51 \pm 0.02	0.55 \pm 0.01	0.60	0.59 \pm 0.02
	Accuracy	0.52 \pm 0.02	0.56 \pm 0.01	0.61	0.59 \pm 0.02

Table 6.4: Results for investigated scenarios - classifiers using cardiac data and acceleration (ACC) as input (from [28]).

Model	Metric	S1	S2	S3	S4
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 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
Random Forest 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 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 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 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 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 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

6.3 Discussion

Capturing emotional events in everyday life poses a challenge, as they occur irregularly and sporadically. Recognizing emotional outbursts with wrist-worn smartwatches and personalized ML models and using this information to trigger self-assessments may improve data collection efforts by increasing the likelihood of capturing emotional states. However, due to the cold start problem, creating such models requires large numbers of per-person training samples. To mitigate it until the necessary number of personal cases is collected, an alternative solution of per-group personalization can be used.

The model adjusted to the group of participants (S3) showed higher classification quality over a general (S1) and a partially adjusted (S2) model. Training models on a large number of general samples with added personal samples (S4) can improve the classification over the general or partially adjusted models (S1 and S2). However, such models usually fail to outperform the adjusted model (Scenario S3). We conclude that, while the quantity of the training set impacts the prediction, its quality and resemblance to the target task were more critical for the model's predictive ability, i.e., we observe better performance for models trained on data from the application domain.

Our results contradict claims that one can easily find emotional cues in human physiology that can be generalized to the whole population. Therefore, all emotion researchers should be wary of the *cold start problem* when designing studies or creating predictive models and try collecting data from new subjects as soon as possible to perform model personalization. Although creating subject-specific models or adjusting them separately for each participant could lead to even better results, our attempts at such a scenario did not yield satisfactory results. We believe that it was caused by insufficient amount of per-subject data, as during the first two weeks of Study B we collected from 13 to 33 (avg 23.7) samples per participant. This low-data problem further supports our claims about sparsity of real-life annotations and difficulty of data-collection endeavors.

Chapter 7

EMOTIONS IN THE WILD

The idea of conducting a large field study (LarField) emerged from our interest in everyday-life human psychophysiology. Starting in 2021, we performed preliminary investigations using a preliminary version of our system for collecting data. We conducted trial studies with a dozen subjects for up to a few weeks to validate specific system components, such as the smartphone application and self-report questionnaires. It led us to formulate the aim and scope of the study in the first quarter of 2022. Its primary objective was to explore if physiological signals collected with wearables can be used to identify physiological patterns of daily-life emotional responses. Moreover, we aimed to research using the same signals for recognition of affect, well-being, and sleep quality.

Contents of this chapter originate from the co-authored article, published in peer-reviewed conference materials:

- [25] J. Komoszyńska, D. Kunc, B. Perz, A. Hebko, P. Kazienko, and S. Saganowski, “Designing and executing a large-scale real-life affective study,” in *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, IEEE, 2024, pp. 505–510.

In the original article [25], our contributions included designing the experiment, collecting data, investigating the results, and drawing the lessons learned, with special contributions of myself to describing the process of data collection and organizational obstacles. Additionally, results described in Sec. 7.3 were created solely for this dissertation.

7.1 Designing a large real-life psychophysiology study

After deciding on the objective, we had to choose the study scope. At this point, it was crucial to find a balance between all the data that could be collected and the data that participants could provide. Our previous research [15] showed that factors such as study complexity and cost, quality of data, participants’ load, and expected compliance rate had to be taken into account. We decided that collecting physiological data and emotional self-reports was the most important for us. However, we

decided to add an additional wearable, together with morning and evening surveys related to sleep, stress, and quality of life, as they provided more information that could prove vital for our research. Also, having more data makes it possible to carry out broader research with the collected dataset. Conducting the study in an iterative form allowed us to refine and improve it based on ongoing observations and to better address financial, organizational, and technological constraints.

Every iteration consisted of approximately 25 subjects tasked with carrying wearable devices and providing self-assessments using the Emognition system [26, 196]. Apart from data collected using the system, each participant filled out pre- and post-study questionnaires, providing additional context, among others, demography, personality, and emotion perception (Tab 7.1). We also collected feedback so we could better understand how participants felt about the study, learn from the user's perspective, and better prepare for future data collection efforts.

7.1.1 Scheduling

After deciding on the study scope, we had to plan functional requirements of Emognition system's, and schedule development of the system and other necessary study components (e.g., measures, consent forms, information for subjects, and ethics committee application). We planned to run pilot studies in the fall of 2022 and start the study in January 2023.

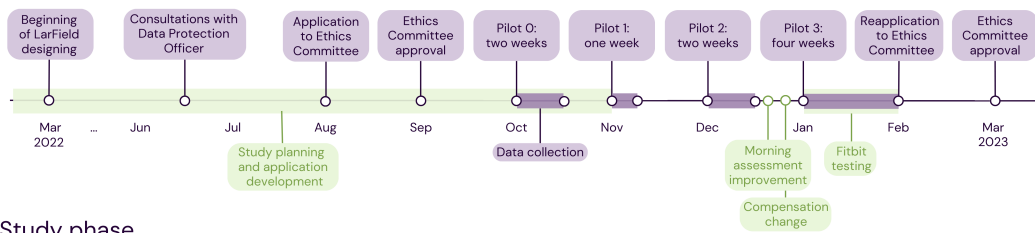
Although we created a schedule with large time buffers between tasks, we still faced substantial delays in development, with some components not being developed at all. Among them, the most critical ones were (1) launching of a tool for monitoring study progress too late, i.e., right before the start of the main study (thus it was not thoroughly tested), and (2) failing to create a tool for validating collected data (both its completeness and quality). Consequences arising from their absence are described in Sec. 7.2.

These delays were caused, among others, by (1) underestimating task complexity and resources needed to complete them (time and people), (2) ignoring the necessity of employing a project manager or, at least, following a project development methodology (e.g., Scrum), (3) unforeseen errors that had to be corrected, and (4) changing the application scope during development. As a consequence, some functionalities, such as collecting audio from the surroundings, had to be abandoned, and study execution (both pilot and main) had to be postponed by two months. We present the resulting timeline in Fig. 7.1.

Table 7.1: Signals and measures collected during the LarField study (from [25]).

Data type	Device	Details
Physiological signals	Samsung Galaxy Watch 3	<ul style="list-style-type: none"> • Photoplethysmography (PPG), Heart rate (HR), RR-interval, Accelerometer (ACC), Gyroscope, Gravity sensor, Ambient light, Atmospheric pressure, Pedometer
	Polar H10	<ul style="list-style-type: none"> • Electrocardiography (ECG), HR, ACC
	Fitbit Charge 5	<ul style="list-style-type: none"> • HR, Resting HR, Heart rate variability (HRV), Pedometer, Distance traveled, Calories burned • (During sleep) Temperature, Respiration rate, SpO2
Contextual data	Smartphone	<ul style="list-style-type: none"> • Approximate location, Events in calendar, Sound intensity, SMS text (Anonymized representation), Phone activity (used applications), Physical activity (detected) • (Metadata) SMS, Calls
Questionnaires during study	Smartphone	<ul style="list-style-type: none"> • (6x a day) Presence of intense emotion, Arousal, Valence, Free text comment • (2x a day) Perceived stress (morning and evening) • (1x a day) Sleep quality (morning), Perceived health (evening)
Pre- and post-study questionnaires	-	<ul style="list-style-type: none"> • (Pre-study) IPIP-BFM-20 [197] (Personality), Demography • (Post-study) Social relations questionnaire, Feedback form • (Pre- and Post-study) SWLS-A [198] (Life satisfaction); SPANE [199] (Affect in life); Flourishing scale [199]; PHQ-4 [200] (Depression and anxiety); PSS [201] (Stress in life); PHQ [202] (General health); RESS-EMA [203] (Emotion regulation); PAQ [204] (Alexithymia)

Designing phase



Study phase

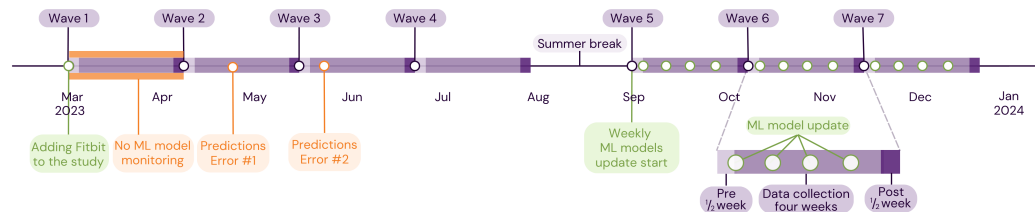


Figure 7.1: Timeline of the LarField phases: designing and main study. From Komoszynska et al. [25].

Assessing the time requirements with more carefulness could have prevented costly delays. Having no prior experience, we could have sought a consultation or employing experienced people, e.g., a technical project manager. Moreover, developed schedule should always be confronted with other responsibilities of team members (e.g., student exams, or teaching duties).

7.1.2 Emognition system

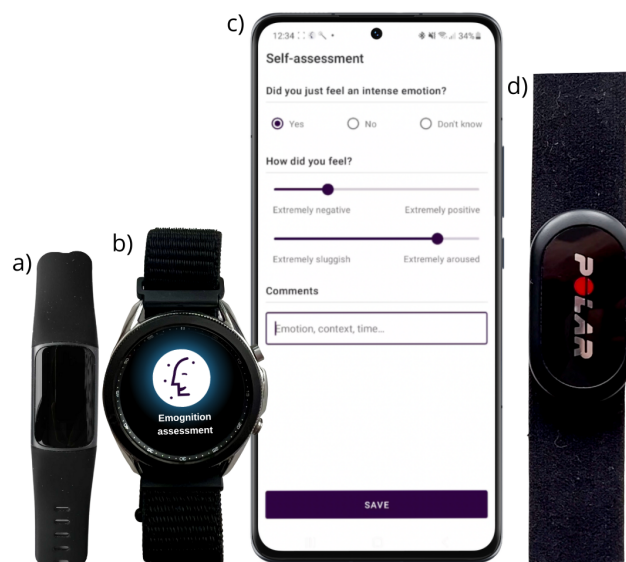


Figure 7.2: Devices used in the LarField study. From Komoszynska et al. [25].

The Emognition system, developed by the Emognition research group from Wrocław University of Science and Technology, was designed to enable conducting real-life emotion psychophysiology studies [15, 26]. It comprises a smartwatch (Samsung Galaxy Watch 3), a chest strap (Polar H10), and a smartphone to gather data using a dedicated mobile application (Fig. 7.2b - d).

The system has gradually developed since 2020. We personally demonstrated one of its early versions at the 2021 Affective Computing and Intelligent Interaction (ACII) conference [15], and it was well-received by the community. After that, the system was further improved to best accommodate our and the subjects' needs [26, 27, 196].

The version of the system used in the study allowed collecting multiple physiological measures from wearables (Tab. 7.1). Smartphone, on the other hand, was used to collect annotations of momentary emotions and daily affect measures, together with contextual data. Additionally, in the LarField study, we employed a smartband that was not integrated with the system (Fitbit Charge 5) to monitor sleep activity (Fig. 7.2a).

The major novelty that the system introduced was utilizing an embedded machine learning (ML) model for detecting intense emotion using real-time physiology. Smartphone application processes physiological data in real-time and uses it for triggering self-assessment questionnaires if it detects an emotional event¹. Such an approach increases the chances of capturing emotional events when compared to random sampling, as it is not precise and may capture only a small subset of emotions that are themselves sparse in real life [27]. Although similar ideas were presented by other authors earlier, e.g., Larradet et al. [205] or Hoemann et al. [206], their detection systems were based solely on heart rate or interbeat intervals, while our system utilizes machine learning models reasoning from multiple physiological signals [27].

7.1.3 Challenges

At the planning stage (June 2022), we contacted the General Data Protection Regulation (GDPR) officer from our university (Wrocław University of Science and Technology) to discuss the data collection process and applicable regulations. The discussion focused mainly on the security and privacy of participants' data during and after the study, and the officer raised some valid points regarding data safety

¹Demonstration video available at <https://www.youtube.com/watch?v=qk3DFmRKKlw>

(e.g., malicious attacks or misuse by data handlers). After researching possible solutions, we chose a process where data was first transferred from subjects' phones to Firebase and later downloaded to the local storage in our laboratory. Although using third-party solutions always raises some concerns, Firebase's reliability, security of protocols used for data transfer, and low vulnerability to malicious attacks exceeded by far any solution that we could develop and host with our resources. Although it was beneficial, the consulting process took longer than we expected, as it involved researching details about data handling and handlers rights, which consumed a lot of resources and time.

Afterward (August 2022), we submitted an application to the Wrocław University of Science and Technology Research Ethics Committee and had to wait for a positive evaluation until September. In January 2023, we submitted an update of an application for the committee's review, as we decided to utilize an additional wearable in our study. After submitting the update, the committee informed us that their regulations do not allow updating approved applications. Therefore, we submitted the second application, which was rejected due to (previously accepted) use of the PHQ-9 [207] questionnaire. As a result, we replaced the PHQ-9 with the PHQ-4, eliminating the objected questions.

Another challenge involved compensation for study participants. Initially, we planned to offer them a choice of either a gift card or a smartband (both worth around 500 PLN). However, because of tax regulations and the university's own rules, neither of those options could be realized. Devising a solution, i.e., preparing contracts with the university and paying subjects in cash, took several months and was inconvenient for study participants and the university's financial department. Also, in some cases, subjects had to visit the university multiple times. Moreover, as we wanted to award additional incentives for active and conscientious subjects (actively engaged in data collection), we had to sign two separate contracts (for base and bonus remuneration), duplicating the required paperwork.

7.2 Conducting a study

Collecting physiological data requires conducting a study. We divided our efforts into three stages: (1) a preparatory stage (described in the previous section), (2) a validation stage for testing our approaches, and (3) an actual study, where data for future experiments were collected. We decided to conduct pilot studies before the actual main study, as we see them as crucial in assessing the readiness and reliability

of developed tools, protocols, and procedures. Additionally, if any problems arise, introducing necessary changes in study components before the main study, e.g., fixing application logic or adjusting notification times, is not as costly, as during the actual experiment (e.g., losing data from the pilot study is acceptable, while losing data from main study decreases the size of collected dataset).

Both types of studies were arranged in a similar pattern. First, study participants were recruited by advertisements in social media and by word of mouth. After signing up for the study via an online form, people who met the requirements were contacted by the study manager and invited for a pre-study visit. During this meeting, participants were introduced to the specifics of the study, procedures, and their responsibilities, and devices used in the study were handed out. Additionally, they completed questionnaires on their personality, demography, well-being, emotion regulation, and alexithymia and filled out the compensation paperwork. After leaving the laboratory, participants were given a few days of time to get used to the system and its specifics. At the beginning of the following week, the actual iteration of the study launched, and from that point onwards, we collected participants' data for four whole weeks (Fig. 7.1). At the end of the collecting period, participants were invited to the laboratory once more for the post-study visit to return the devices, fill out closing questionnaires, and complete the required paperwork.

7.2.1 Pilot studies

In each pilot study, we validated different components of the study together with possible issues caused by introduced changes, from the recruitment procedure to the final post-study meeting (Tab. 7.2). Predefining each pilot's goals allowed us to focus on specific procedures or tasks, which helped in thoroughly verifying and refining them.

The significance of pilot studies lies in their ability to identify issues that may not surface during regular evaluation. These studies, which involved one internal pilot and three external pilots, were crucial in testing all parts of the study before its main launch. Following each study, a person managing the experiment collected detailed feedback on the Emognition system, study components, and overall experiment experience. Collected feedback, together with frequently asked questions, revealed necessary adjustments in different aspects of the study.

Improving pre-study visit procedures involved refining explanations and additional information materials provided to participants, such as (1) the contents of the pre-

Table 7.2: Main aspects validated in pilot studies. Based on Komoszynska et al. [25].

Pilot ID	No. of subjects	Duration [weeks]	Main aspects validated
0 (internal)	5	2	<ul style="list-style-type: none"> • A data collection system • Private data transfer to a study phone
1	5	1	<ul style="list-style-type: none"> • Procedures: <ul style="list-style-type: none"> – preparing devices – pre-study visit – explaining study to subjects
2	7	2	<ul style="list-style-type: none"> • ML model • Verifying subjects' engagement throughout the study duration • Verifying compensation level • Fitbit testing
3	7	4	<ul style="list-style-type: none"> • Changes in the mobile application

sentation about the study, (2) the mobile application manual, and (3) the study participation consent form. Other adjustments included (4) adding an item return checklist, as some people forgot to bring all of the provided devices back for the meeting, ending the iteration, and (5) adding pregnancy and obesity as exclusion factors, as suggested by the device manufacturer. Also, the Frequently Asked Questions (FAQ) section in the mobile application was improved.

During the study, participants had to fill out two types of questionnaires measuring (1) well-being (on mornings and evenings) and (2) emotional state (several times a day). Initially, we wanted to trigger the morning questionnaire using Android's built-in sleep detection feature. However, the internal pilot showed that this feature did not work as expected. Instead, we let participants provide their sleeping habits and triggered questionnaires based on the provided time. To address possible missed notifications or irregular circadian rhythms, which some participants showed, we updated the application logic to allow completing the well-being questionnaire at any moment during the predefined timeframe.

When planning the study, we assumed that (1) wearables would be worn throughout the day and charged at night, and (2) participants would be prompted at most six times during the 10 hours of recording session (battery life of the smartwatch running our application). However, during pilot studies, some participants did not complete any assessments for several days in a row. It allowed us to identify faulty logic

in our application arising from previously assumed typical daily routines. As we expected participants to use the Emognition system during the day, we scheduled application counters to reset at night. If a person had an active session during nighttime, either because they were awake or forgot to turn off devices, the system identified the session as ongoing and valid. It did not refresh the assessments counter for the day. After consulting the development team about the issue, it turned out that changing this behavior would require a significant application logic update. Therefore, we decided to keep using the same scheduling logic. However, we (1) instructed participants about scheduling and turning off the devices before midnight and (2) added a notification alerting participants if any wearable remained operating after 11 p.m. Those measures significantly reduced the problem.

The Emognition system collects not only physiological signals and corresponding assessments but also contextual data, including approximate location, used applications, and data on calls and SMS (without raw content). To avoid issues with the system's stability on different devices, we provided participants with all hardware necessary for the study, i.e., all wearables and a smartphone. We asked subjects to migrate to provided phones and use them as their primary (or only) smartphones throughout the study. Aware that such change may be hard for some subjects, as it required them to transfer all of their data to the provided phone, we decided to pilot this approach after much careful consideration. As team members were not objective in this matter, it became one of the most important elements to be tested during external pilot studies.

Most study participants did not object to phone change, with some pleased with the possibility of using a phone newer than their own. However, pilots revealed that not all people comply with the requirement, with some of them swapping SIM cards between two phones, using two phones simultaneously, or not migrating to the provided phone at all. These observations led to implementing a SIM card monitoring feature, which told us if the phone was used as intended and allowed us to intervene on such occasions. Our interventions considerably reduced undesired participants' behavior.

Pilot studies also disclosed that much time was required to prepare devices for participants. We optimized this process by creating scripts to automate it partially. It ensured that all devices were set up in the exact same way (settings, application version) and considerably shortened the preparation time.

The above examples show that pilot-testing the study is crucial, as it enables iden-

tifying issues beforehand and reducing the risk of failure during the main study. Improvements in procedures and components made them less prone to subjects' undertaking unexpected actions. We believe that piloting could be further improved by making some participants intentionally misbehave, try to break the system or fool the procedures.

7.2.2 Main study

We conducted the main LarField study from March to December 2023, running a total of seven iterations. Although we pilot-tested study components and procedures, some significant issues occurred. The development of the data quality and completeness monitoring tool was delayed, and we had to launch the study without it. Because of that, we could not see if the study was proceeding without issues, if participants were wearing devices and collecting data as instructed, and if all modules of the Emognition system were working as intended. Although monitoring could be done manually by screening the database, it would be too time-consuming to assess 25 participants, given other tasks. It was not until we finished the first wave and examined their data that we discovered incorrect functioning of the emotional self-assessment triggering using the ML model, a key system function. It resulted in the first wave's emotion self-assessments triggered solely at random moments. The issue was caused by incorrect logic in adjusting thresholds for ML model output and setting their level to values that were too high. As it occurred during the main study, and we did not have enough time to test any major changes thoroughly, thresholds were adjusted manually in all following iterations.

As we could not monitor the study data in real-time, two other problems with the ML model's triggering mechanism remained undetected until later in the study (second and third wave, Fig. 7.1). First, we discovered that the data transfer rate from watch to phone was insufficient in some cases and only when specific conditions were met, e.g., large size of data waiting in a transfer queue. Once detected, it was easily fixed by increasing the analyzed data buffer size. The second issue was harder to detect and also involved data transfers. Occasionally, no data could be transferred between the smartwatch and the phone, which was caused by the library used for transferring files. Attempts at fixing this issue failed, and the only solution found to work involved restarting both devices.

Later, once we improved the study monitoring tool, we could respond to the problems on the fly. For instance, when we noticed unusual data and error reports from one

subject, we investigated the issue with them. It turned out they performed a factory reset of the smartphone and tried to handle the situation on their own rather than reporting this to our team. The participant installed an old version of the Emognition application, which could be found online but was no longer used by us. As a result of this lesson, we put more stress on asking participants to report any issues or concerns related to the devices or system to us. We also removed the old version of the application from the internet.

Another tools that should had been developed earlier, but were not until the end of the study, were scripts for checking completeness of data and cleaning it. It held back the analyses, which could not start until the tools were completed. Because of that, we discovered late that data about calendar events were not collected properly. We wanted to use it as contextual data for annotated emotions, but it was implemented in a too simple way. To quicken the implementation of this feature, we gathered information only about new events added using the phone. It ignored changes in events, events added using other devices, and those created before the study. If discovered earlier, this feature could have been improved. Also, more precise requirements for developers could help, as they could plan longer for the implementation and test if it works as intended.

The inability to analyze the data resulted in other information being improperly collected, i.e., text messages and call tracking. The Emognition applications tracked only the built-in communication applications and ignored external applications popular among some participants, especially younger ones. Fortunately, our application collects usage data about other applications so that some limited communication metadata can be obtained.

7.3 Results

During the seven months of the study, 167 participants took part in the experiment (86 females, M age = 29.7; SD = 10.2, 77 males, M age 26.5, SD = 8.6; 4 other or not specified, M age = 27.0, SD = 8.8). Almost all participants were Polish, with two of Belarusian origin and one of Ukrainian. Additionally, one participant identified as Polish-Moroccan. Interestingly, two Poles declared Polish not to be their mother tongue.

The study population consisted mainly of university students (100, 60%). Additionally, 74 participants had already completed a degree of some level. Employed participants most often declared their job to be in education (23, 13.8%), science

or engineering (12, 7.2%), public administration and services (10, 6.0%), general administration (7, 4.2%), culture (5, 3.0%), or service activity (5, 3.0%). Almost all participants saw their financial situation as similar (71) or better (73) than an average person's. Most subjects (105, 63%) declared being in a relationship, and 29 people (non-students only) had children (M no. children 1.9, SD = 0.6).

We collected one month's worth of everyday psychophysiological data from each study subject. This effort resulted in more than 35,000 hours of physiology recordings and over 20,000 emotion self-reports, accompanied by contextual data from additional questionnaires and a smartphone, adding up to 2.5TB of data. Although we have not finished processing physiological data at the time of writing, we managed to perform a brief analysis of emotion reports and a few simple experiments on the dataset that we presented below.

In emotion and daily states recognition experiments, we treated each problem as regression and classification (binary, high/low value) tasks. We used four machine learning algorithms to solve them, i.e., K-Nearest-Neighbours (KNN), Multilayer Perceptron (MLP), Random Forest, and Support Vector Machine (SVM). We report model accuracy in terms of Root Mean Square Error (RMSE) for the regression task and F1 score for classification. As in some cases, the majority class constituted even 87% of annotations, we decided to use macro-averaged version F1 metric, which is well-suited for experiments with both balanced and imbalanced data.

Additionally, we examined the feasibility of personalization by comparing three approaches, namely: (1) training models only on preceding assessments (from previous day for morning states, from the same day for evening states) and physiology (utilized features were the same as in one of our articles [28]), (2) training models on preceding data, with additional context of personality scores for each of dimensions from personality questionnaire, and (3) training models on preceding data with additional context from personality and demography questionnaires. We call those models General, Personality, and Broad context, respectively. Although none of the strategies used an entirely independent set of participants, we treat approach (1) as producing models close in performance to general ones because it contains much less personal information than two other approaches. These approaches were examined in tasks of recognizing daily morning and evening affective states and emotions. In Sec. C.1, we present questionnaires that were utilized to gather affective state ground truth and personal information used as context.

The results are compared with respective baseline predictions from an average

annotation predictor in regression and a majority class predictor in classification tasks. For obtaining model performance estimates, we followed the procedure designed by Bouckaert and Frank [208] using 10-fold cross-validation repeated 10 times, together with corrected paired t-test for determining statistical significance of differences: (1) between model results and baselines, and (2) between different approaches to model design on within-condition averaged results (results averaged across all four algorithms for each task and condition, e.g., general models for predicting evening health). All results are reported as averages from different runs with standard deviations for both models and baselines. All p-values were corrected using the Holm–Bonferroni procedure [209] (where applicable).

7.3.1 Personality and emotions

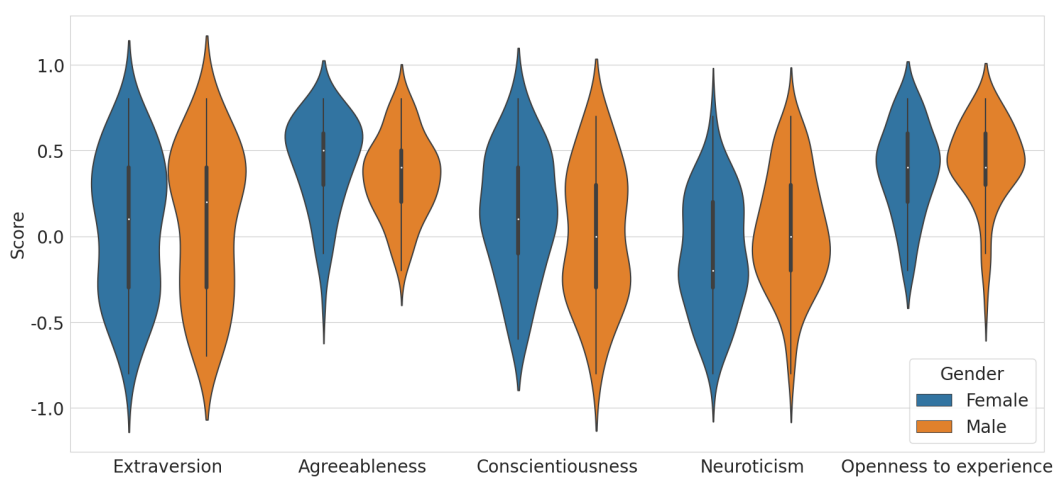


Figure 7.3: Distributions of Big Five personality model scores in LarField dataset.

The first analysis focused on examining distributions of personality scores in the dataset and their relationship with emotion measures. Violin plots of collected scores for the Big Five personality model (Fig. 7.3) show some differences between study subjects. Distributions of extraversion and neuroticism cover the whole scale of values (-1 to 1) for both male and female participants. Additionally, in male participants, we observed the same behavior in conscientiousness. Moreover, women scored on average (median) higher than men in agreeableness and conscientiousness, and men scored higher in extraversion and neuroticism. These observations are supported by kernel density estimations, with men tending to score around the average or higher in extraversion and neuroticism, and female scores showing the same behavior in agreeableness and conscientiousness. Despite no visible difference in median openness to experience scores, they tended to be around the 0.5

Table 7.3: Results for emotion measure regression from personality trait scores.

Gender	Personality trait	Emotion measure	r^2	p
Female	Agreeableness	Arousal	0.000	0.963
		Valence	0.020	0.199
	Conscientiousness	Arousal	0.033	0.099
		Valence	0.020	0.194
	Extraversion	Arousal	0.013	0.293
		Valence	0.104	0.003
	Neuroticism	Arousal	0.051	0.038
		Valence	0.126	0.001
	Openness to experience	Arousal	0.001	0.730
		Valence	0.016	0.244
Male	Agreeableness	Arousal	0.002	0.712
		Valence	0.062	0.032
	Conscientiousness	Arousal	0.295	0.000
		Valence	0.083	0.012
	Extraversion	Arousal	0.048	0.058
		Valence	0.121	0.002
	Neuroticism	Arousal	0.127	0.002
		Valence	0.204	0.000
	Openness to experience	Arousal	0.056	0.041
		Valence	0.053	0.046

value or higher for men and more spread out between 1 and -0.5 for women.

To examine the relationship between personality traits and emotion measures, we performed a simple linear regression by taking average personality scores for each person and trying to model person-averaged arousal and valence annotations. It revealed a few small but significant² linear relationships (Tab. 7.3) between personality traits and emotion measures, with five of them at the $p < 0.05$ level, four at $p < 0.01$, and two at $p < 0.001$. The highest value of the coefficient of determination (r^2) was at 0.295 for arousal regression from conscientiousness, and the lowest among the significant ones was at 0.053 for valence regression from openness to experience, both in men. Also, we noticed that such regression was more often significant in men than in women, with eight out of ten experiments achieving statistical significance in men and only three in women. These results led us to suspect that models for affect recognition on this dataset might benefit from personalization.

²Coefficients of determination (r^2) and p-values from Wald Test with t-distribution for null hypothesis that the slope is zero were computed using `linregress` function from SciPy package [190].

7.3.2 Morning and evening state recognition

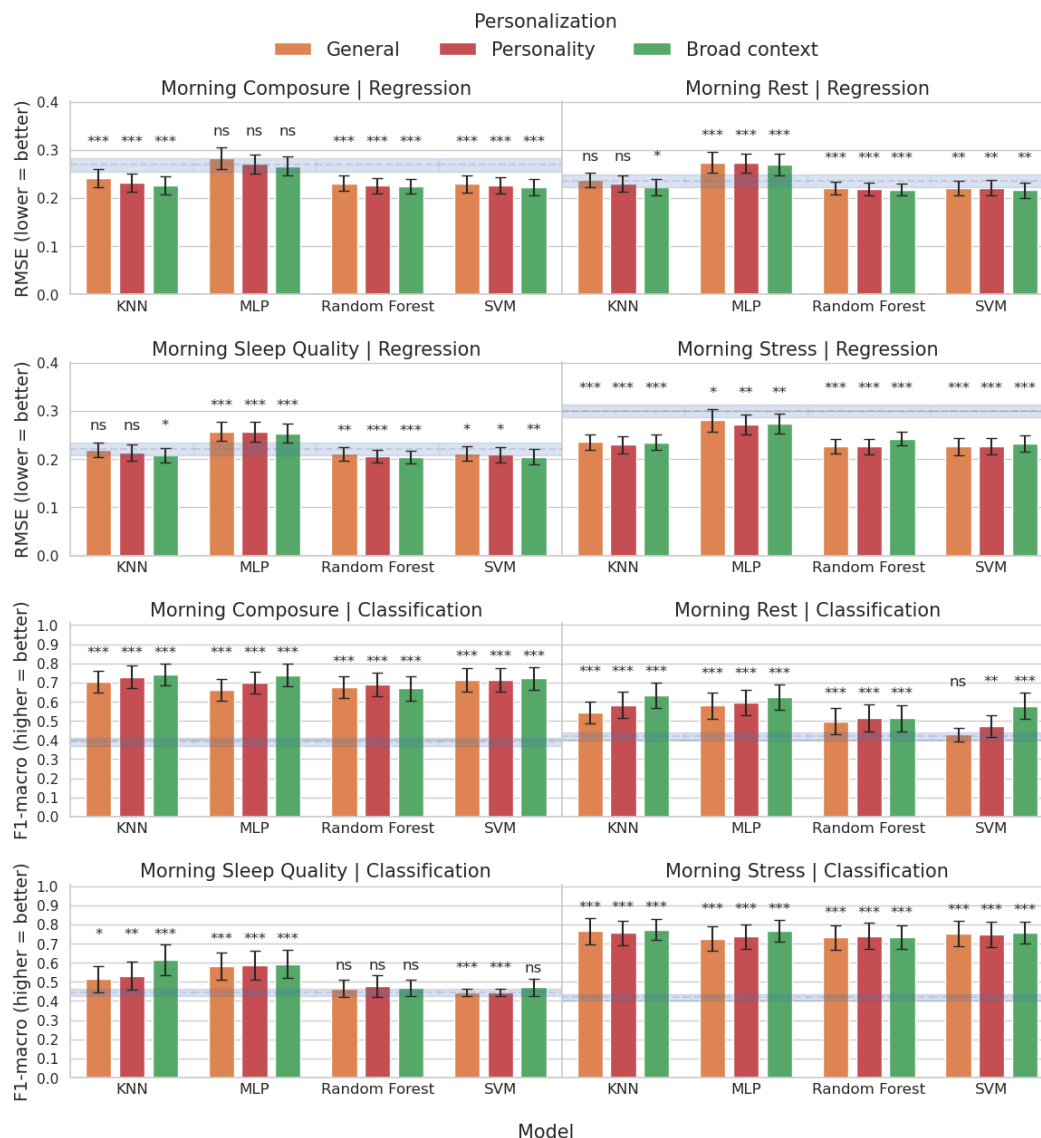


Figure 7.4: Results of morning questionnaire prediction on Larfield dataset. Average (regression) or majority (classification) annotation baselines (values from test sets, horizontal line denotes average baseline prediction \pm standard deviation). Statistical significance between models and baselines from corrected paired t-test [208]; ns - not significant, * - $p < 0.05$, ** - $p < 0.01$, *** - $p < 0.001$.

In the first batch of experiments, we tried predicting daily states that can be regarded as indicators of self-reported stress, mood, and health, namely: (1) composure, stress, rest, and sleep quality measures from morning questionnaires, and (2) health, mood, overwhelm, and unpredictability measures from evening assessments.

In experiments where morning self-assessments were predicted, most of the cre-

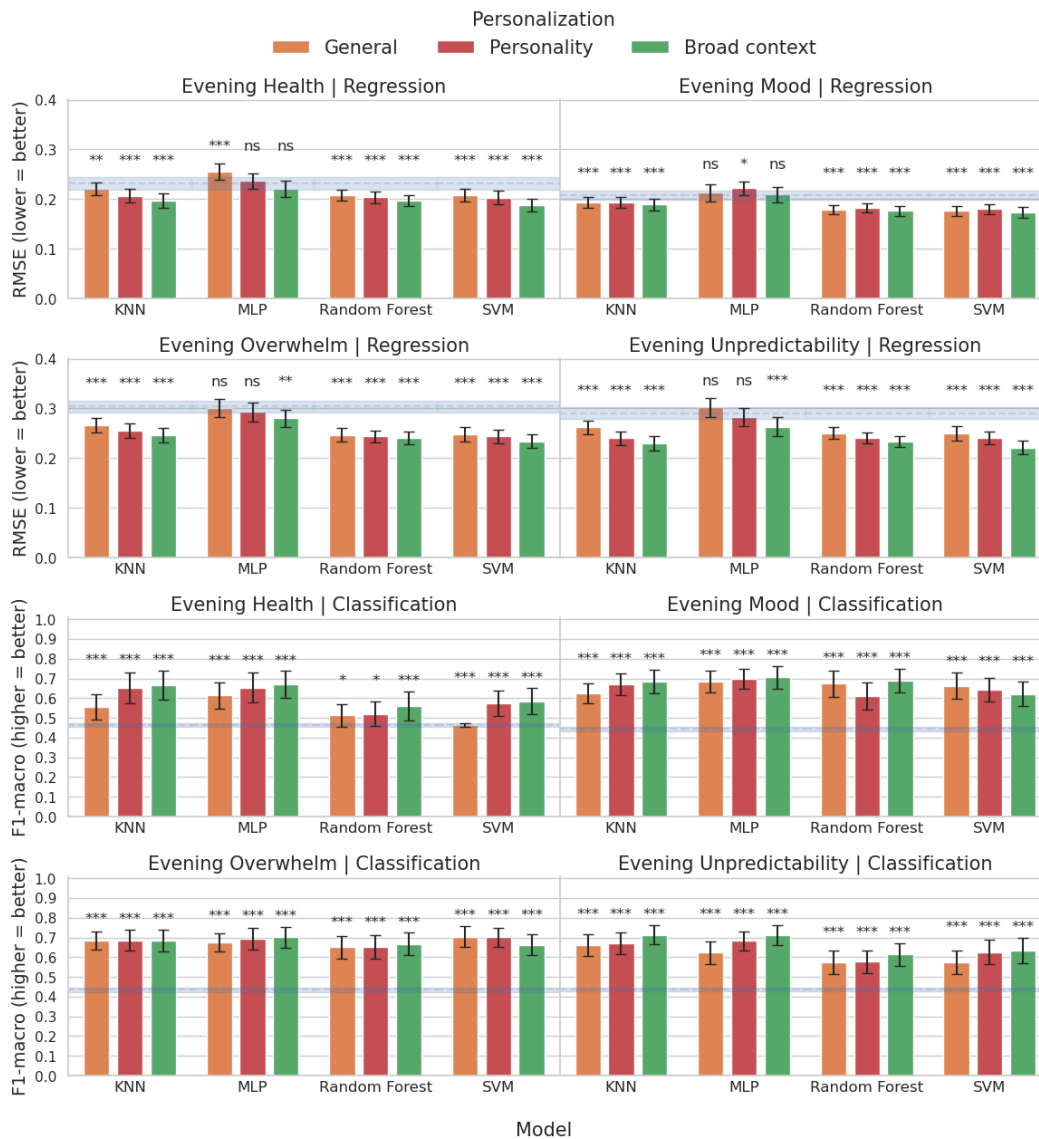


Figure 7.5: Results of evening questionnaire prediction on Larfield dataset. Average (regression) or majority (classification) annotation baselines (values from test sets, horizontal line denotes average baseline prediction baseline prediction \pm standard deviation). Statistical significance between models and baselines from corrected paired t-test [208]; ns - not significant, * - $p < 0.05$, ** - $p < 0.01$, *** - $p < 0.001$.

ated models achieved significantly better results than respective baselines (Fig. 7.4, Tabs. C.1 - C.8). The only models that performed significantly worse than baselines were MLP in morning rest and sleep quality regression. Out of the classification models, most achieved accuracies significantly better than baselines, except for general SVM in rest, broad-context SVM in sleep quality, and all Random Forest models in sleep quality classification. Morning stress recognition turned out to be relatively

simple to predict, with all regression and classification models performing significantly better than baseline predictors, and classification models achieving overall highest values of F1-score among all morning states.

Experiments with predicting evening questionnaire responses yielded similar results as morning experiments (Fig. 7.5, Tabs. C.9 - C.16). In each task, two MLP models for regression performed significantly worse than baselines. Overall, models for predicting mood were among the best-performing ones in terms of F1-macro and achieved the lowest RMSE.

In most tasks, both in morning and evening state recognition, we noticed that including personal context improves results over the general models. Pairwise comparisons of modeling strategies for morning assessment predictions (Tabs. C.34 and C.35) showed that differences in achieved prediction accuracy were significant only between General and both types of personalized models in composure regression. In the classification of morning states, significant differences were present between general and broad-context models in rest and sleep quality recognition. In both of those tasks, broad-context models performed significantly better than personality-aware ones. In rest classification, even including just personality context improved results over general models. Models for regression of self-reported evening states (Tabs. C.36 and C.37) were less homogeneous. In the regression task, only general and personality-aware models for mood recognition showed no significant differences, suggesting that in all other tasks, personalized models performed significantly better than general ones. In evening health and unpredictability classification, both personalization strategies resulted in significantly higher F1 scores than the general approach.

7.3.3 Daily emotion recognition

After modeling morning and evening assessments, we decided also to try recognizing answers to emotion assessments. First, we grouped emotion questionnaires, based on the time of their creation, into three groups, namely (1) morning questionnaires – filled after the morning assessment and noon, (2) afternoon questionnaires – filled between noon and 6 p.m., and (3) evening questionnaires – filled between 6 p.m. and evening assessment.

In experiments where morning self-assessments were predicted, most of the created models achieved significantly better results than respective baselines in classification and significantly worse results in regression (Figs. 7.6 and 7.7, Tabs. C.17 - C.28). In

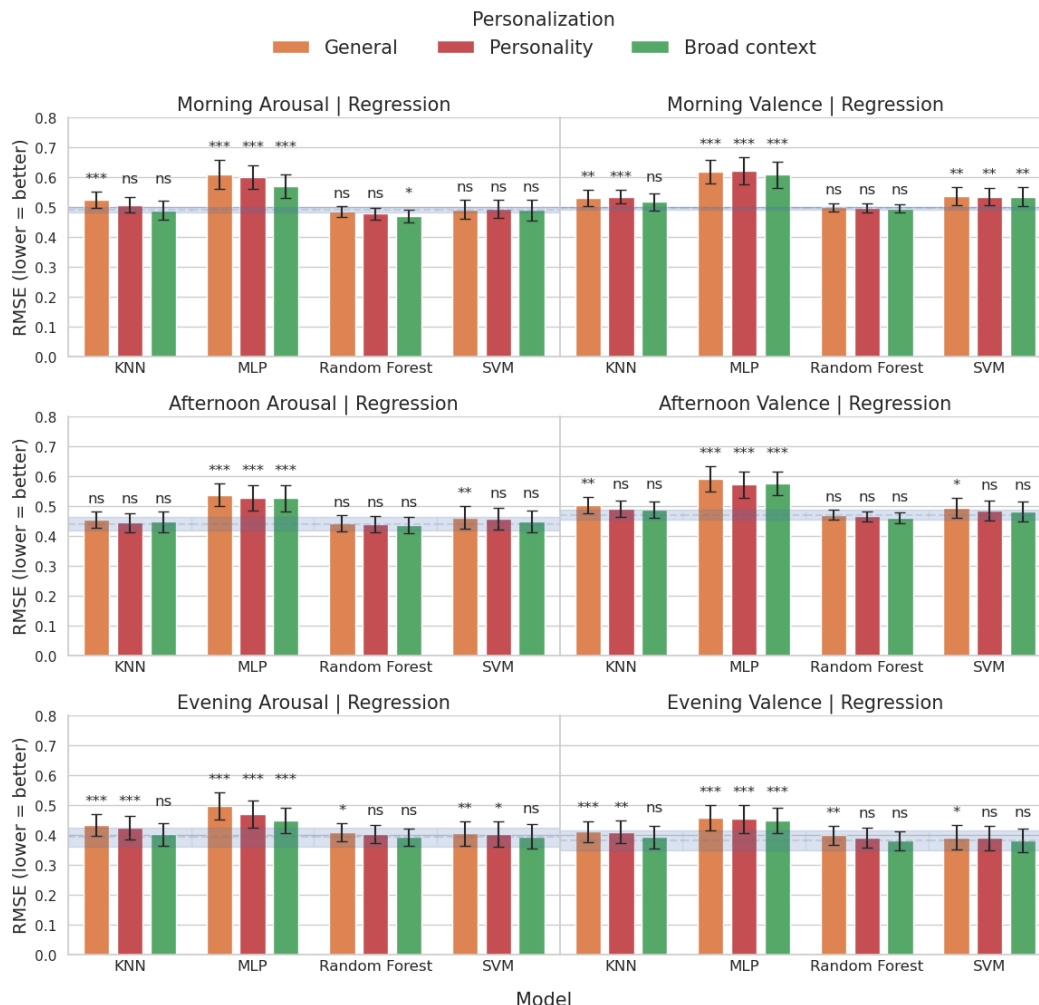


Figure 7.6: Results of emotion questionnaires regression on Larfield dataset in daily setup. Average annotation baselines (values from test sets, horizontal line denotes average baseline prediction \pm standard deviation). Statistical significance between models and baselines from corrected paired t-test [208]; ns - not significant, * - $p < 0.05$, ** - $p < 0.01$, *** - $p < 0.001$.

all regression tasks, MLP models performed significantly worse than baselines, and the only model outperforming the average baseline was the broad-context Random Forest. On the other hand, in classification models, most models achieved accuracies significantly better than baselines, with broad-context models performing on a baseline level only in afternoon arousal recognition (Random Forest and SVM). In daily emotion recognition, morning arousal can be regarded as the simplest to predict, with all classification models outperforming the baseline and broad-context Random Forest achieving error rates significantly lower than the respective baseline as the only regression model.

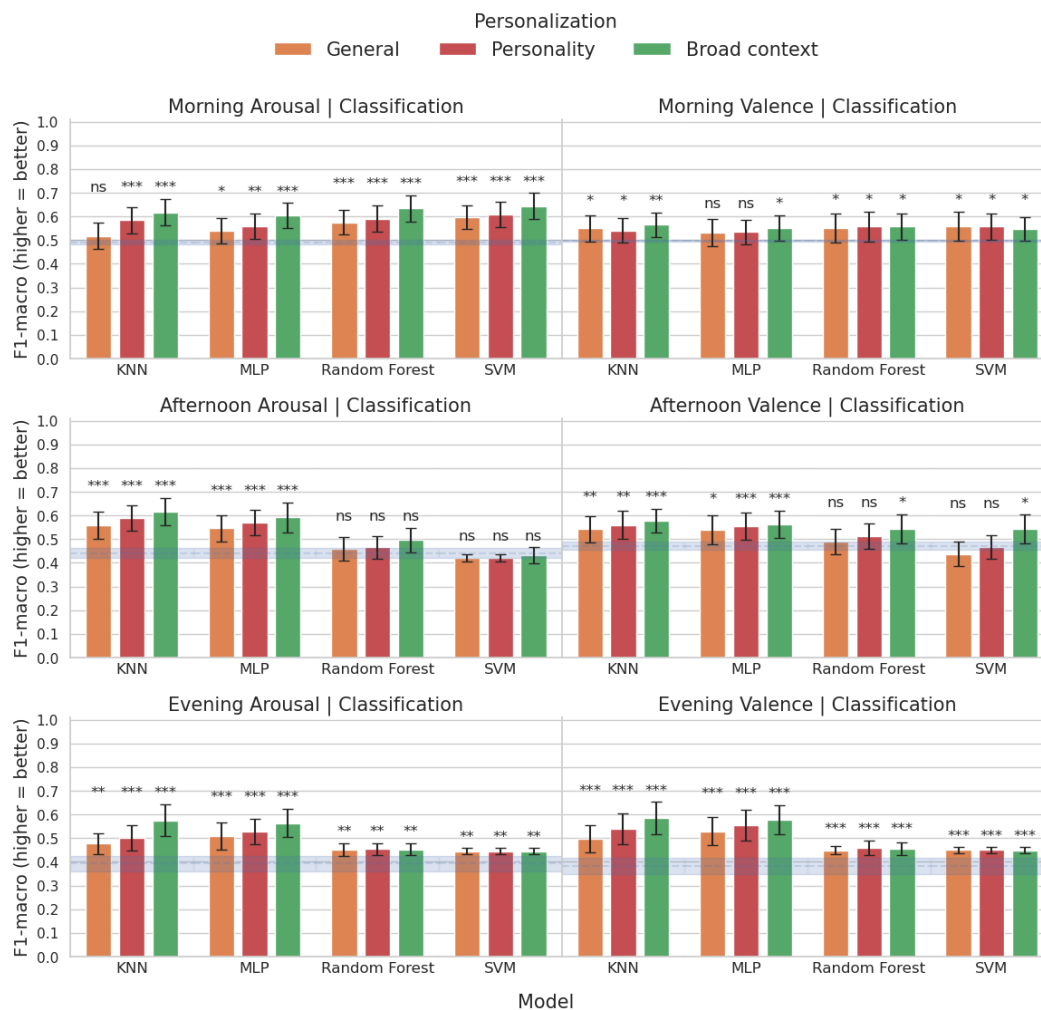


Figure 7.7: Results of emotion questionnaires classification on Larfield dataset in daily setup. Majority annotation baselines (values from test sets, horizontal line denotes average baseline prediction \pm standard deviation). Statistical significance between models and baselines from corrected paired t-test [208]; ns - not significant, * - $p < 0.05$, ** - $p < 0.01$, *** - $p < 0.001$.

Pairwise comparisons of personalization strategies for daily emotion assessment predictions (Tabs. C.38 and C.39) showed significant differences in achieved error rates between broad-context and two other methods in morning and evening arousal regression, and between personality and general approach in evening arousal. In classification, models using broad context performed on the same level as general models in morning valence and on the same level as personality models in morning and evening valence, outperforming them in every other task. Personality-aware models for classification showed significantly better results than general ones only in morning arousal classification.

7.3.4 Momentary emotion recognition

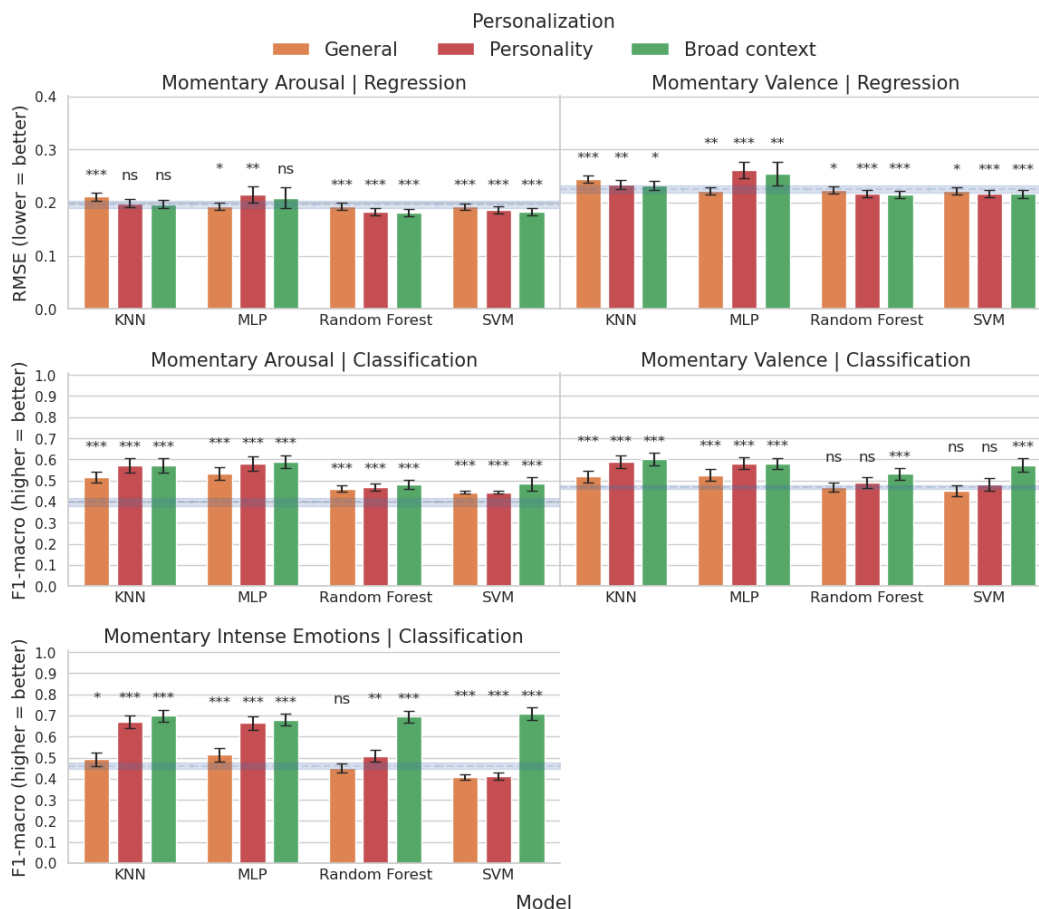


Figure 7.8: Results of emotion questionnaires prediction on Larfield dataset in momentary setup. Average (regression) or majority (classification) annotation base-lines (values from test sets, horizontal line denotes average baseline prediction \pm standard deviation). Statistical significance between models and baselines from corrected paired t-test [208]; ns - not significant, * - $p < 0.05$, ** - $p < 0.01$, *** - $p < 0.001$.

The last batch of experiments on the LarField dataset focused on recognizing self-reported momentary emotions from momentary physiology, morning assessment, and contextual information depending on the modeling strategy. We attempted to recognize arousal and valence as classification and regression tasks, and the occurrence of intense emotion, being itself a binary state, as a classification task only.

In experiments where momentary self-assessments were predicted as regression tasks, most of the created models achieved significantly worse results than the average baseline (Fig. 7.8, Tabs. C.29 and C.30). In regression tasks, only Random

Forest and SVM models performed significantly better than baselines, achieving lower RMSE in personalized training. On the other hand, MLP models for regression performed better than baselines only when trained using a general approach. Pairwise comparisons of three modeling strategies showed that, on average, there are no differences between them in momentary emotion regression (Tab. C.40).

In classification, on the other hand, most models performed significantly better than respective baselines (Fig. 7.8, Tabs. C.31 - C.33). On average, in all three tasks, broad-context models achieved the highest F1-macro scores, followed by personality-aware models, which was confirmed by pairwise comparisons of modeling strategies (Tab. C.41).

7.4 Discussion

In this chapter, we described the performed study, aimed at collecting a dataset of real-life physiology and behavioral information annotated with momentary emotions, complemented by demography and contextual data from daily morning and evening questionnaires. This dataset may be used for testing novel methods for real-life emotion recognition and for uncovering patterns in outside-of-laboratory affective states. Its large population and amount of per-participant data allow it to be used to test hypotheses about both general and personalized models for emotion recognition.

We share the description of study preparation, issues that we encountered, and lessons that we learned, so other scientists may learn from our effort, and improve on our methods. We also performed a few simple experiments to better understand the data that we collected and tested hypotheses that we hoped the dataset would allow us to examine. Unfortunately, as at the time of writing the dataset is still being processed, and first experiments were possible as late as in June 2024, we could perform only simple experiments on data that did not require precise synchronization of different devices before the thesis submission date.

7.4.1 Designing real-life emotion psychophysiology study

Most of the challenges that we encountered were caused by delays in the study preparation and could have been avoided if (1) we had taken appropriate measures against them or (2) the study launch had been postponed until all components and tools were coded. However, the timeline specified in the funding grant did not allow us to delay the study any longer. To prevent such situations, one could employ a technical project manager to look after the development process and ensure the

Table 7.4: Selected challenges encountered during the design and execution of the study, along with our recommendations in handling them (from [25]).

Challenges	Recommendations / possible solutions
Managing the development of the study system and procedures	Consider appointing a project manager or applying a project management methodology, which could improve the academic style of project management.
Overseeing the study	Hire a study manager responsible mainly for the study (designing, monitoring) and participants (recruitment, training, supporting).
Interacting with external entities, e.g., with an ethical committee	Incorporate a large time buffer for tasks associated with external entities, for example, submitting the application to an ethical committee way in advance.
Planning tests	Plan in advance what and how has to be validated. In the case of pilot studies, perform at least two; determine their goals, and allocate time for making corrections.
Validating devices	When choosing devices for the study, analyze them in person. Pay attention to data availability, data quality in various conditions, battery life, ease of usage, and convenience. Be open to replacing the original component, e.g., a smartwatch strap, to improve adjustment and fit.
Advanced research equipment and participants' visit procedures	Create a step-by-step protocol, specify personnel's responsibilities, and implement automation where feasible.
Last minute changes in the study, e.g., adding devices or questionnaires	Carefully assess the potential advantages and drawbacks of changes. Consult the change with all stakeholders. Remember to update the ethical approval.
Making the study bulletproof	Conduct several pilot studies and collect feedback. Simulate a noncompliant participant. Monitor participants and data during the study to be able to respond to any detected anomalies. Introduce monitoring of variables that can potentially have impact on the study results.

deadlines are met. On the other hand, the traditional workflow in academia rarely assumes such supervision so that it could become a source of conflicts within a team.

Employing a study manager with a background in psychology was the right decision on our end. Employing one three months before the planned pilot studies allowed them to grasp the study concept and learn all the study-related procedures. This allowed them to efficiently handle the recruitment process (advertising the study, contacting and training participants) and administrative duties (discussing agreements, contracts, and consents), attend to pre- and post-study visits, monitor study progress, and complete other minor tasks. Moreover, it allowed other team members to focus on developing applications used during the study and solving major issues as they arose.

We also learned that all decisions related to data-collecting devices (wearables) should be assumed early in the study planning process. After careful deliberations about their benefits and drawbacks, we decided to utilize Fitbit smartbands in the study. However, it turned out that we had to abandon integrating them into the data collection system, as it would introduce additional costs and possible bugs due to late changes. As a result, we had to configure smartbands and create user accounts manually and could not monitor data collected using Fitbit devices in real time.

Considering all of the constraints in the affective study development process requires meticulous work and attention to its intricacies. Besides following standard procedures, researchers must also consider the ethical risks that such a study may induce and try to minimize them. Fortunately, some researchers are focusing on examining these aspects and recommending possible solutions to them, thus simplifying organizing such experiments [69].

7.4.2 Affective state modeling

Results of the conducted investigation suggest some interactive effects of personal features (e.g., gender or personality) on real-life affect. A simple regression of per-person averaged arousal and valence scores as a function of a personality trait showed non-negligible effects, especially in neuroticism, extraversion, and conscientiousness in men and neuroticism in women.

Albeit using only simple daily features, such as morning or evening mood, stress and health questionnaires, average daily physiology, and daily annotations of emotions, outperforming the baseline in predicting morning and evening affect was possible

in both classification and regression tasks. Also, in the regression of morning composure and all evening states but rest, and in the classification of morning rest and sleep quality, along with evening health and unpredictability, we noticed significantly better results in models that utilized personal information and the ones that did not.

Emotion recognition experiments were conducted in two setups – recognition of averaged daily states and of momentary emotions. Both setups showed that regression of emotions is a difficult problem. In daily emotion tasks, personalization resulted in a significant improvement of error rates only in arousal prediction, and in momentary emotions, there were no significant differences between the three strategies. On the other hand, in classification, models using broad context achieved significantly higher F1 scores than general ones in all tasks but morning valence. Using only personality as context improved results over no context in morning arousal and all momentary emotion tasks.

Overall, while we noticed some potential for improving emotion recognition by including personal features, achieved gains were highly dependent on the task. Our results showed that while predicting daily morning and evening self-reported affect was in general easier (models often outperformed baselines), there were often no significant differences between personalized and general modeling strategies. In emotions, which were seemingly harder to predict, we can notice improvements from the additional personal context in classification but almost no improvement in regression.

Our results are limited by several factors. First, due to the insufficient amount of time, caused by delays in the study and data processing, created models were relatively simple and therefore had low capacity. Additionally, no optimization was performed for the utilized models, as after removing data samples that could not be used in these experiments, we were left with a relatively small dataset. Allocating the additional validation subset was impossible in this case, making the whole setup prone to potential overfitting caused by optimization. Moreover, annotations themselves may not be as precise as would be required in regression, causing confusion during training and leading to imprecise predictions.

These factors could impact the results in several ways, either benefiting general or personalized models. Low model capacity and lack of parameter optimization may have increased the difficulty of learning general patterns with a very limited set of features or made it impossible to learn patterns from a wide variety of useful

features as the volume of the feature space increased. Moreover, all of the predicted annotations were subjective, thus making it hard to find general patterns in data, and also limiting the amount of data that could be used when modeling specific people or groups. The fact that emotion perception is itself impacted by emotions and that some participants reported in their feedback that the study allowed them to gain more insight into everyday emotions, which they rarely think about, adds to the claim about the overall subjectivity of reported data. The overall better performance of models for binary classification, when compared with regression models, could be then caused by the task's nature, which focuses on coarse estimates rather than precise measures of emotional states.

*Chapter 8***SEARCHING FOR PHYSIOLOGICAL MARKERS OF EMOTION**

Up to this point, this dissertation focused on reviewing existing literature, describing our efforts in collecting datasets for researching the psychophysiology of emotions, and recognizing those states. However, although it is well-known that emotions and human physiology are somehow connected, there is still no reliable evidence for the presence of unique markers of emotion in human physiology. Motivated by this problem, we organized an Emotion Physiology and Experience Collaboration (EPiC) challenge. Its main point was to evaluate the strengths and limitations of using machine learning to model presumed connections between peripheral nervous system (PNS) activity and self-reported emotions. The challenge was held as a part of a workshop organized at the 2023 Affective Computing and Intelligent Interaction (ACII) conference. To ensure replicability and transparency, we designed the challenge with three goals in mind:

1. Evaluating how well can researchers use PNS features to model emotion using one openly-available dataset.
2. Studying the impact that validation approach choice has on prediction quality.
3. Promoting open science by making all resources publicly available.

In this chapter, we describe the competition structure and obtained results and summarize our findings. We commit to promoting open science(the third item) by making all resources publicly available in a public Open Science Framework (OSF) project¹.

Contents of this chapter originate from the co-authored article², undergoing a review in a peer-reviewed journal at the time of writing:

¹OSF Project "2023 Emotion Physiology and Experience Collaboration (EPiC) Challenge" is available under the link <https://osf.io/bmhsd/>

²Article preprint is available in the OSF project repository.

- [32] N. A. Coles, B. Perz, M. Behnke, J. C. Eichstaedt, S.-H. Kim, T. N. Vu, C. Raman, J. Tejada, G. Zhang, T. Cui, S. Podder, R. Chavda, S. Pandey, A. Upadhyay, J. I. Padilla-Buritica, C. J. Barrera Causil, L. Ji, F. Dollack, K. Kiyokawa, H. Liu, M. Perusquia-Hernandez, H. Uchiyama, X. Wei, H. Cao, Z. Yang, A. Iancarelli, K. McVeigh, Y. Wang, I. M. Berwian, J. C. Chiu, M. Dan-Mircea, E. C. Nook, H. I. Vartiainen, C. Whiting, Y. Won Cho, S.-M. Chow, Z. F. Fisher, Y. Li, X. Xiong, Y. Shen, E. Tagliazucchi, L. Bugnon, R. Ospina, N. M. Bruno, T. A. D’Amelio, F. Zamberlan, L. R. Mercado Diaz, J. O. Pinzon-Arenas, H. F. Posada-Quintero, M. Bilalpur, S. Hinduja, F. Marmolejo-Ramos, S. Canavan, L. Jivnani, and S. Saganowski, “Big team science reveals promises and limitations of machine learning efforts to model the physiological basis of affective experience,” *Nature Human Behaviour*, 2024, In reviews.

In the project [32], our contributions involved designing the challenge and methodology used, conducting the challenge, investigating the results, and drawing conclusions, with special contributions of myself to designing testing procedures and drawing conclusions regarding methods and models utilized by competition participants, curating the dataset, validating participating teams’ code and results, supervising the technical aspects of the challenge, and describing modeling approaches and rationale of validation methods.

8.1 Competition as research method

Emotional experience is a complex process in which physiological, psychological, and cognitive components intertwine. Although many researchers have been studying the impact that emotions have on PNS responses, the scientific community remains unconvinced by results presented thus far (Chap. 3, 4). The reasons for such state of affairs include (but are not limited to) (1) complexity of studied states, (2) lack of interdisciplinary knowledge within research groups, (3) inability to replicate and validate presented evidence, and (4) limited resources.

Although it is impossible to address all of the problems the affective-research community faces, some of them can be researched using a big team science approach [210, 211]. For example, examining the impact of emotions on bodily reactions is a complex task that a unidisciplinary team could not possibly solve. It requires knowledge of psychology, physiology, signal processing, and computer science, which are too broad for one researcher or a small research group to handle. In such situations, big team science allows everyone’s expertise to be used to perform experiments and draw conclusions. Although different teams could not cooperate

during the competition, we can study similarities and differences in their approaches and results to draw conclusions.

8.2 Materials and methods

Taking all of our objectives into account, we decided to use the Continuously Annotated Signals of Emotion (CASE) dataset [212] in the challenge. We divided the data into four validation scenarios, each of them highlighting different dependencies in the data, and invited 15 teams to participate in the challenge.

8.2.1 Dataset

The original dataset consists of data from 30 subjects (15 females, M age = 25.7 years, SD = 3.10; 15 males, M age = 28.6 years, SD = 4.8), who were exposed to a total of 8 emotionally stimulating videos, targeting states of amusement, fear, boredom, and relaxation (two videos for each emotion). While watching stimuli, participants provided continuous ratings of their emotional experience using a joystick (in valence and arousal space). Additionally, their PNS activity was measured during all inductions using eight physiological signals: blood volume pulse (BVP); electrocardiography (ECG); electromyography (EMG) over the zygomaticus major, corrugator supercillii, and trapezius muscles; electrodermal activity (EDA); respiration (RSP); and skin temperature (SKT).

In the CASE dataset [212], subjects and stimuli have their own specific identifiers. For the competition, we decided to obfuscate data by assigning each subject and video a different random number drawn independently for each validation scenario (see next subsection). Randomization was performed once for each scenario, so in scenarios utilizing cross-validation subject and video numbers were consistent between folds.

8.2.2 Training and testing procedure

The challenge consisted of predicting valence and arousal ratings from a person's physiological signals. We designed four validation scenarios to evaluate prediction quality, each emphasizing different interrelations in data. Out of them, one scenario was designed as a hold-out validation (one data split into training and test sets) and the rest as cross-validation (multiple splits of the same data into training and test sets)³. Additionally, in every test set, the emotion self-reports were 20 seconds

³We did not provide separate validation subsets to participants, but we permitted using training data as they pleased. We also allowed them three early submissions that were evaluated on half of

shorter than the physiology recordings⁴, to allow modeling momentary emotions using signals preceding or following each of them. The scenarios are illustrated in Fig. 8.1 and described below.

1. The **across-time scenario** used a *hold-out validation* approach and focused on the order of signals in time. We divided each recording in the dataset (each induction for each person) into training and test sets based on time. The first part of the recording was included in the training set, and the later part was included in the test set. This scenario can be seen as testing the usefulness of knowledge regarding a person's reaction obtained from past data for predicting new reactions in similar situations.
2. The **across-subject scenario** used a *leave-N-subjects-out validation* approach. Data were randomly divided into five groups (folds), so every fold consisted of all the observations from six people. In every cross-testing pass, such fold was used either in training (multiple times) or test set (only once). Conceptually, this scenario examined the generalization abilities of models created for one group of people when predicting emotional ratings provided by a different group of people.
3. The **across-emotion scenario** used a *leave-one-emotion-out validation* approach. As mentioned before, videos in the experiment targeted four emotions: amusement, fear, boredom, and relaxation. Thus, to test how well models generalize between conditions (targeted emotions), we created four folds, each consisting of different inductions (videos of one kind).
4. The **across-induction scenario** used a *hold-out validation* approach. As a reminder, during the experiment, the same emotion was targeted twice, using two different videos (e.g., two inducing fear). Each pair of videos was split into two folds - data from one induction in the first fold and data from the other in the second fold. This scenario examined the quality of emotion rating prediction by models trained on a different induction of that same emotion.

The above scenarios were designed to allow modeling emotional experience using physiology in a replicable and easy-to-understand manner. They also provide a good framework for testing models for possible overfitting.

the test data.

⁴Exactly 10s of unannotated physiological signals both before and after annotations.

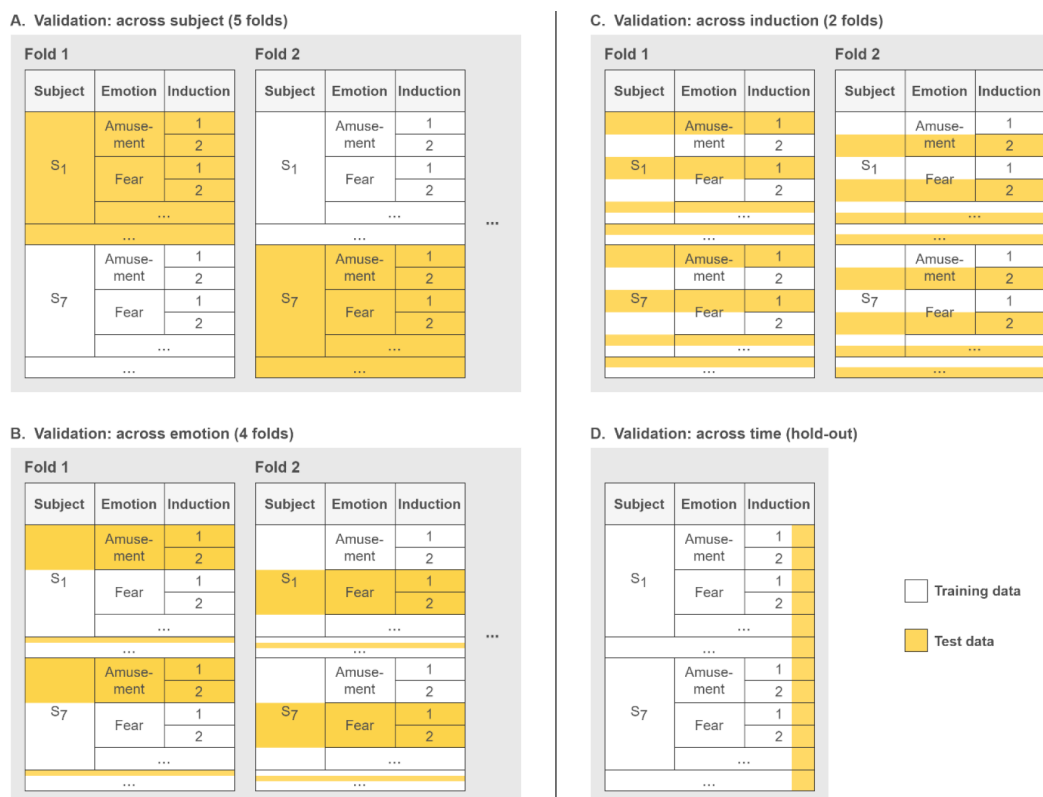


Figure 8.1: Validation scenarios employed in EPiC competition (from [32]).

To prepare the challenge data, we had to divide it into training and test parts. Training data always began at the beginning of the original recording, and its length varied depending on the duration of the original induction. For each stimulus, the length of corresponding training samples was chosen once per scenario, so they have a constant duration within a scenario. Additionally, we ensured that the duration of created learning cases differs by, at most, a few seconds between scenarios. This varying length was used to further obfuscate subject-induction pairs between different scenarios.

Test cases were created similarly to training cases, with instances coming from one induction within the validation scenario coming from the same part of the recording. Following a standard approach to validation, all inductions for all subjects were used for testing exactly once within a scenario. As chronological order was vital in the across-time validation scenario, we split each session from the CASE dataset (subject-video pair) into three fragments – training data (beginning at the start of the session), buffer data (discarded part of data lasting around the 20s), and test data (starting after the buffer). Other validation schemes assigned the whole session either to training or validation sets, so there was no possibility for training and

testing on overlapping samples. Therefore, we did not introduce the buffer in other scenarios, allowing training and validation data to contain common samples.

All test cases in the competition had the same length and structure. They consisted of 50-second-long physiological recordings and centered 30s of corresponding annotations. The unlabeled fragments of 10s at the beginning and end of each test case were introduced to allow using short periods of past and future physiological data (relative to annotation moment) for reasoning. Participants were provided with test cases where physiology remained untouched, but all annotations were deleted, except for their timestamps.

We employed root-mean-square-error (RMSE) as a metric for evaluating machine learning models during the competition and arranging the leaderboard. Additionally, we used absolute prediction error and two baseline models for further analyses regarding the quality of submitted predictions and their suitability for drawing conclusions about physiological markers of emotions. The used baseline models were (1) a random prediction model, assigning a random value to each point in the test set, and (2) a mean prediction model, assigning an average computed over all annotations from the training set to each point in the test set (done separately for each data split).

Moreover, to better understand obtained predictions and analyze results even further, we (1) reviewed the code of top-performing submissions and (2) ran three of them, deemed most rigorous in their approach, on samples where physiology was substituted with random noise. For an arbitrary data split, let us denote a test dataset of size N , consisting of physiology recordings X_i and their annotations Y_i , as $T = \{(X_i, Y_i) : i \in 1..N\}$. For this analysis, we created another dataset, where each physiology recording X_i was replaced with random time series (white Gaussian noise, $M = 0.0$, $SD = 1.0$), therefore creating $\tilde{T} = \{(\tilde{X}_i, Y_i) : i \in 1..N\}$. We then ran models trained for the challenge on the altered dataset (\tilde{T}) and compared predictions between samples from T and \tilde{T} .

8.2.3 Teams and modeling approaches

Out of 18 teams that applied for the challenge, S.S., M.B., and N.C. selected teams based on their (1) expertise in the field, (2) experience with similar challenges, and (3) a proposed approach. To be eligible to participate, teams had to agree to (a) refrain from cheating (e.g., using an original dataset for model development), (b) write their code in an accessible way and make it openly available, and (c) collaborate

on a paper summarizing the challenge. The code was inspected for the three top-performing teams (and a subset of other teams) to ensure that it was accessible, replicable, and contained no evidence of cheating. To motivate participants, we awarded \$300 to all teams that completed the challenge and an additional three \$200 performance awards based on the overall prediction quality, i.e., root mean squared error (RMSE) across all validation scenarios. Available funds allowed us to recruit up to 15 teams for the competition, of which 12 teams completed it and three resigned during the challenge. Below, we present a brief description of each team's modeling approach⁵ (more details can be found in Tab. D.1).

Team 1 (AffectiveBulls) Physiological signals were used to derive a univariate time-series representation through a weighted sum of all signals, with normalized variance used as weights. This signal was split into windows of fixed duration and used for training person-specific Neural Networks that predict valence and arousal dimensions at the same time.

Team 2 (Cafeteros) Physiological measures were cleaned and used to obtain derivative time series, such as phasic and tonic components of EDA. Using those signals, the team trained hybrid deep learning models consisting of convolutional and recurrent branches (TCN-LSTM). For each data split, they trained one model for predicting arousal and one for valence.

Team 3 (CARElab) Physiological signals were cleaned, split into windows, and used to derive descriptive features. Next, machine learning models were trained for the narrowest possible context in each scenario using AutoGluon [213], a machine learning framework for automated model training and optimization.

Team 4 (UBA) The team trained tree-based ensemble models (random forest and gradient-boosted decision trees) for predicting emotional state in the narrowest possible context, using features derived from preprocessed physiology as input.

Team 5 (PSU) Physiological signals were cleaned and used to compute dynamic features. These features were used as an input for trained machine learning models - deep transformer models or ensembles of decision trees.

⁵Code submitted by participants and competition details can be found in the challenge GitHub repository: <https://github.com/Emognition/EPiC-2023-competition>

- Team 6 (Northeastern)** Physiology was first cleaned, and descriptive features were derived. The team trained several different classical machine learning regression models and chose the best-performing one for each scenario and affect dimension. They predicted annotations for every 1s and upsampled them later.
- Team 7 (NYIT)** The team used a model taking raw physiology as input, which was scaled only. They employed their own version of FEDformer [214], a deep learning transformer-like architecture for time series forecasting. The same model predicted arousal and valence levels, which were later smoothed using an additional convolutional deep learning model.
- Team 8 (Princeton)** For scenario 1, they used signal windows with annotations centered relative to physiology and an LSTM model with information about the subject and video. For other scenarios, they used signal windows with annotations at the end and regularized LSTM models working on physiology provided for the challenge. They predicted arousal and valence simultaneously.
- Team 9 (Queens)** Preprocessed ECG signal (time-series) was used to train deep neural networks: convolutional and transformer-based architectures. The authors explored two approaches: (1) training models from scratch and (2) pretraining models on the entire training data and retraining them in a narrower context. Out of the tested approaches, the transformer model trained from scratch on the narrowest possible context achieved the best accuracy.
- Team 10 (SAIL)** They cleaned and resampled physiology. Skin temperature signal was used in the form of features, and the rest of the signals were used as time series. For modeling, they used a deep state-space S4 model and transfer learning, where they trained layers for ECG signal processing on another, much larger dataset. They trained separate models for predicting arousal and valence.
- Team 11 (IITB)** The team first cleaned and resampled physiology, which was later split into 2s windows, with annotations at the end of each window. Using this data, they trained decision trees to predict arousal and valence scores (separately).

Team 12 (VSL) Signals from the provided dataset were utilized as input for transformer-based models. Input data was normalized and scaled inside the model and later processed using neural network layers. They trained models to predict valence and arousal at the same time.

Out of the above, only four teams created models generating arousal and valence predictions at the same time (teams 1, 7, 8, 12). Seven teams focused solely on creating deep learning models (1, 2, 7, 8, 9, 10, 12), three teams solely on classical machine learning methods such as tree-based or SVM models (4, 6, 11), and two tested various approaches and selected the best-performing ones for each scenario (3, 5).

We can also categorize models employed by teams based on their approach to the context present in the data in each validation scenario. Three teams (3, 4, 9) always trained their models in the narrowest possible context in each, meaning that they utilized induction-specific models in across subject scenario (A), subject-specific models in across emotion (B), and across induction (C) scenarios, and distinct models for each subject-induction pair in across time (D) validation scenario. Team one approached the competition in a similar way but created one general model for predicting emotion experience ratings in across emotion scenario (B). In across time scenario (D), team 8 also utilized scenario-specific knowledge and provided their model with information about the subject and induction from which each sample originated. Other teams trained general models, predicting core affect ratings for the whole population. Additionally, the approaches of the four teams did not follow the strict validation procedure employed in the competition, possibly creating overfitted models (teams 1, 5, 7, 11)⁶.

8.3 Results

To examine each team's prediction quality, we used mixed-effect regression. We regressed each team's absolute prediction error as a function of (1) the prediction source, i.e., the team's model or one of the baselines (random or the average values), (2) the scenario (validation approach), (3) a higher-order interaction between model source and scenario, and (4) random intercepts for each subject and stimulus. We used random intercepts to include non-independent observations in the analysis (originating from one person-video pair). For each validation scenario, we used

⁶Information in this paragraph were mainly derived from analyzing code submitted by participating teams, as most of them did not provide detailed enough descriptions of their approaches.

pairwise differences in regression coefficients (contrasts) to estimate the mean difference (MD) in the absolute prediction error between each team's model baseline model and test its significance (Fig. 8.2). Results show that all teams who completed the challenge performed better than a random baseline model ($1.89 < MD > 0.48$, all $z > 130.31$, all $p < .001$). However, the mean baseline model was harder to outperform, with only seven teams (58%) achieving it in across-subject ($0.49 < MD > 0.01$, all $z > 2.44$, all $p < .05$) and across-time validation scenarios ($0.73 < MD > 0.02$, all $z > 5.13$, all $p < .001$); five teams (42%) in across-emotion scenario ($0.70 < MD > 0.02$, all $z > 5.81$, all $p < .001$); and three (25%) teams in across-induction scenario ($0.51 < MD > 0.23$, all $z > 62.58$, all $p < .001$).

We can also observe that prediction accuracy did not behave consistently between scenarios but varied between teams (Fig. 8.2). For instance, team 1 prediction error was lower in across-subject vs. across-time validation, while team 4 results showed the opposite pattern.

These results highlight the multiplicative nature of constraints on generalizability, i.e., interactive effect on inferences of decisions about (a) used models and (b) utilized validation.

The above results can be seen as preliminary evidence for links between PNS activity and core affect reports. However, the fact that ML models outperformed the mean baseline model only in some cases raises questions about the nature of their predictions. At this stage, it is unclear whether they were driven by potential links between people's PNS activity and core affect reports or simple averages of annotations observed in the training set.

To investigate the emotion-PNS activity relationship further, we performed additional experiments on data with input samples substituted with random noise, the logic being that if ML models do not capture emotions and PNS activity, we should see no significant difference between prediction results. We tested it using mixed-effect regression and regressing each team's absolute prediction error as a function of (1) the scenario in question (validation approach), (2) whether the prediction came from real or random data, (3) a higher-order interaction between the scenario and whether the input was real or simulated, and (4) random intercepts for each person and stimulus in the dataset. In 93% of tests, the teams' prediction quality decreased for models tested on noise input when compared with the same models tested on real input ($-0.33 < MD > -0.02$, all $z < -6.56$, all $p < .001$), with only one case showing no significant difference ($MD = 0.00$, $z = -1.01$, $p = .31$) (Fig. 8.3).

These results provide further evidence that some of the submitted models utilized PNS signals when predicting self-reported emotions.

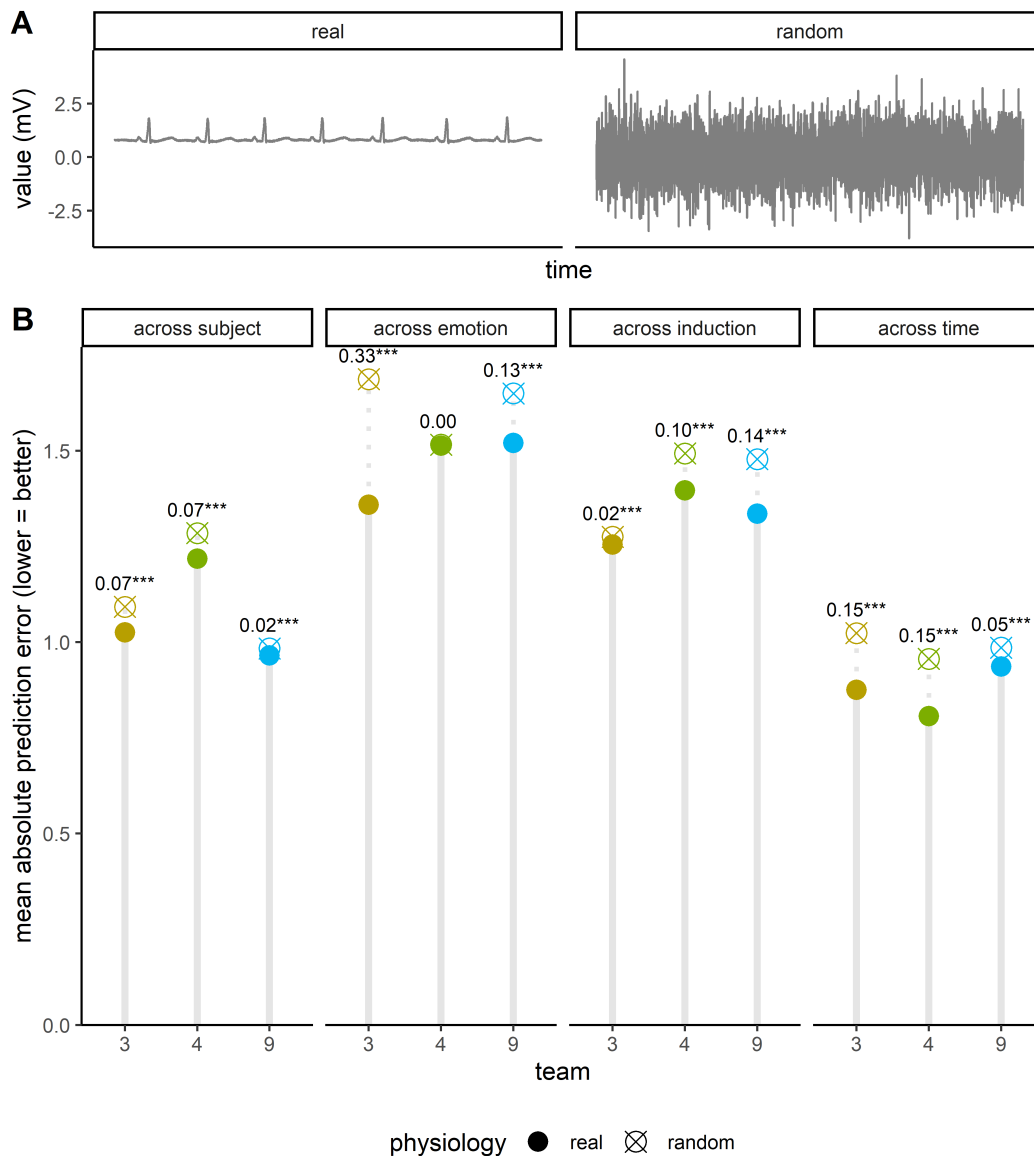


Figure 8.3: Panel A – example of real and simulated random electrocardiography signal. Panel B – mean absolute prediction errors (y-axis) for predictions made by models created by 3 teams selected for additional testing, reported for four validation scenarios. Results are reported together with mean differences (***) denotes $p < .001$). Circle – results when tested on real physiology; crossed circle – results when tested on simulated physiological randomness (from [32]). For specific values, see Tab. D.6.

8.4 Discussion

The above results show that machine learning can potentially model complex physiological patterns associated with emotional experience. Models created by 12 teams of researchers all performed better than random guessing, and about half of them also outperformed the average baseline created from training annotations.

An interesting observation arises when comparing employed approaches to creating models and competition results. It turns out that the top three teams (out of those that strictly followed validation procedures) all used models trained within the smallest possible context – induction-specific models in across subject scenario (A), subject-specific models in across emotion (B), and across induction (C) scenarios, and subject-induction specific models in across time (D) scenario. Those results show that in the employed dataset, psychophysiological dependencies were best modeled by often small but specific models rather than the ones learning general patterns.

Additional tests with simulated physiological randomness add to this conclusion, as it caused a drop in prediction accuracy in nearly all models from three selected teams. However, in many tests, the differences in prediction quality between real and random physiology were small, demonstrating the need for improvements in utilized modeling strategies.

8.4.1 Emotion psychophysiology

Obtained results are in line with theories that posit a physiological basis of emotional experience [46, 215], and at the same time with competing theories postulating spuriousness of these links, e.g., that both PNS activity and emotions are independent, and both originate from upstream neural process [216]. However, this work was not designed to provide strong tests regarding causal relationships between PNS activity and emotion, e.g., team 4's model [217] indicated that using preceding (vs. succeeding or co-occurring) physiological signals to predict self-reported emotion yielded the best accuracy. This observation could support the claims about causality in the emotion-forming process [218], but such patterns may originate from delays in participants' emotion reports.

Our results highlight commensurability and generalizability challenges in emotion research utilizing machine learning [219]. The results show that differences in study design, i.e., how models are developed, benchmarked, and tested, can impact final conclusions. In our case, focusing on a random baseline (vs. mean baseline) or across-time validation (vs., for example, across-emotion validation) would result

in an overly optimistic image of the models' accuracy. These results contribute to an ongoing discussion about emotion recognition technologies and their potential benefits (e.g., unobtrusive measurement of emotions) and harms (e.g., the possibility of inaccurate predictions) [17, 18].

We also provide a more challenging observation that constraints on generalizability can be multiplicative (i.e., interactive) [220]. Our results show that obtained accuracy depended both on the modeling and validation approaches without a clear distinction of a more influential factor. Thus, we conclude that they had an interactive effect on results. This observation leads to the conclusion that before making theoretical claims about emotions, researchers should carefully study the data and the assumed experimental design. For example, team 1 models achieved better accuracy (lower prediction error) in across subject vs. across time scenario. It could be seen as evidence for greater within-person than between-person variability of links between physiology and emotion reports, leading to claims for biologically-innate [221] nature of emotion. Meanwhile, results of other teams (e.g., team 8) show the opposite pattern. They could be regarded as evidence for high inter-subject variability and against biological innateness of associations between PNS activity and self-reported emotions [222]. Similarly, observations that context-specific models performed in general better than general ones may suggest that we should focus on developing small but tailored solutions for emotion recognition. However, this dependency may be specific to the utilized dataset, and no conclusions should be drawn before testing the phenomenon on more datasets using appropriate procedures.

8.4.2 Big team science

This work also serves as a proof-of-concept study for using big team science [210, 211] in affective research. By standardizing data sources and testing procedures, we obtained cleaner comparisons of different approaches to the same problem. By making variations in methodological decisions systematic (e.g., employing multiple testing procedures), we could examine how these decisions impact the generalizability of inferences. Also, requiring that competitors make their code openly available allowed us to further inspect and reproduce teams' solutions and identify the most promising approaches for future research [223]. As all materials were published, they may be useful for other researchers working on similar problems. Overall, big team science effectively allowed us to leverage the wisdom of crowds to evaluate a challenging theoretical question in emotion research.

Our examination also yields lessons about conducting collaborative research and how it can be improved. For example, competition organizers could not thoroughly examine all submissions' code, nor could they comprehensively evaluate the behavior of all models. The task could be crowdsourced [224], but it would require establishing protocols and best practices for code review. Also, many teams pointed to insufficient resources as an issue, such as a need for (1) more time to create models, (2) larger datasets, and (3) high-performance computing resources. Recent initiatives regarding dataset development [210, 211] and providing shared computing resources [225] may help overcome these limitations. Despite these constraints, our work presents a possibility of utilizing collaborative efforts to address complex questions in science with the help of novel machine learning methods. However, researchers must remember about the multiplicative constraints on generalizability when applying these methods.

PERSONALIZED DATA PROCESSING

In the previous chapter, we explored group-personalized models and showed their potential in creating models for emotion recognition and in mitigating the cold start problem. In this chapter, we delve deeper into the data preparation process and explore how different processing methods impact the results of personalized models, both for groups of people and individual subjects. We apply those methods to both input signals and annotations and thus name them a *two-fold personalization*. Moreover, based on our research described thus far, we compare different experimental setups (general / group-personalized / individual models) to see which of them performs best on the used datasets in terms of absolute accuracy metrics and achieved gain over baseline models.

9.1 Materials and methods

9.1.1 Datasets

In the research, we utilized four datasets well-established in affective computing, namely *A Dataset for Affect, Personality and Mood Research on Individuals and Groups* (AMIGOS) [226], *a multimodal database for implicit personality and affect recognition* (ASCERTAIN) [227], *Continuously Annotated Signals of Emotion* (CASE) [212], and *a Database for Emotion Recognition through EEG and ECG Signals from Wireless Low-cost Off-the-Shelf Devices* (DREAMER) [228]. We chose them as all of them: (1) contain electrocardiography (ECG) physiological measures, which we decided to focus on in this research; (2) contain emotions annotated using a multivalued dimensional arousal-valence model, allowing us to perform both classification and regression tasks, (3) are highly cited, between 160 and 820 times¹. Moreover, each of them has distinct characteristics that may allow us to explore different aspects of emotion recognition, with CASE being annotated in a continuous manner and AMIGOS, ASCERTAIN, and DREAMER being collected using relatively unsophisticated measuring devices resembling everyday-life wearables.

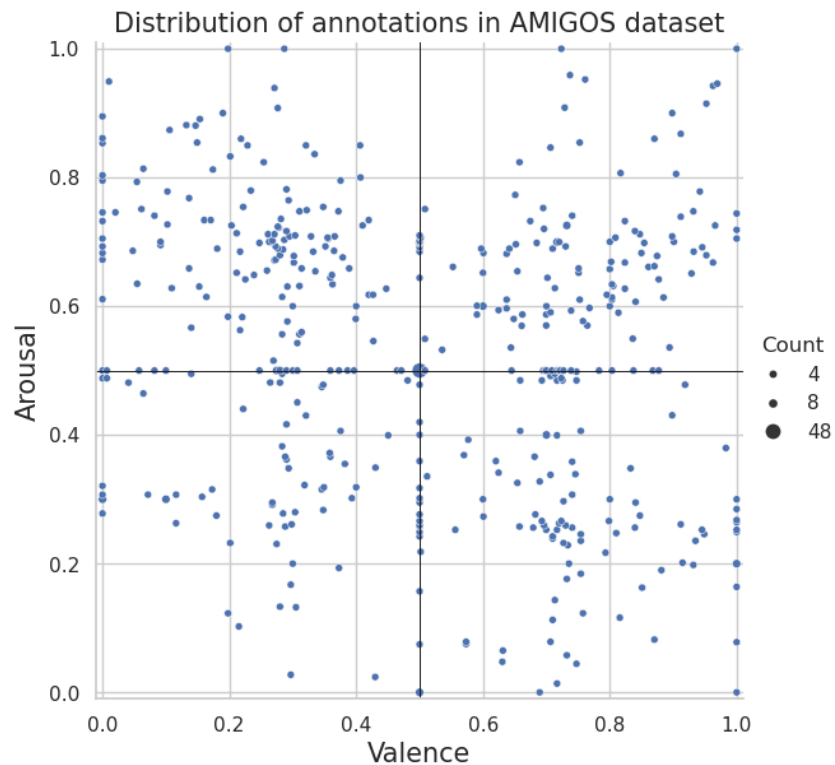


Figure 9.1: Distribution of annotations in a subset of AMIGOS [226] dataset used in experiments. All values were normalized relative to ranges of scales in employed questionnaires.

AMIGOS

The original dataset consists of data from 40 healthy participants (13 females; M age = 28.3). All subjects were exposed to 16 short video stimuli, each targeting a specific quadrant of the two-dimensional arousal-valence plane (4 videos per quadrant). Participants provided their emotions after watching stimuli, using 9-point scales for rating their perceived arousal, valence, dominance, liking, and familiarity, and binary selection for basic emotions of neutral state, disgust, happiness, surprise, anger, fear, and sadness. During all inductions participants had their PNS activity measured using electrocardiography, electrodermal activity, electroencephalography (14 electrodes), and video data. All PNS measures are collected using simplified interfaces registering signals using only a couple of electrodes (two for ECG, one for EDA, and 14 for EEG). Although used devices are not suitable for use in everyday life studies, they are not fully laboratory-grade either, thus making this dataset a step towards real-life emotion research. Authors also provide scores from *Positive and*

¹According to Google Scholar search engine <https://scholar.google.com/>

Negative Affect Schedule (PANAS) questionnaire and personality scores for the Big 5 Personality model (extraversion, agreeableness, conscientiousness, neuroticism, openness) for each participant.

When exploring the dataset, we noticed that some of the registered signals were of very low quality. We decided to exclude recordings showing poor quality in the last 35 seconds of an ECG signal. For automated scoring of signal quality, we utilized the Neurokit 2 [229] Python toolbox. Additionally, we entirely excluded ten participants who had less than ten such samples. The resulting distribution of emotion annotations is presented in Fig. 9.1.

ASCERTAIN

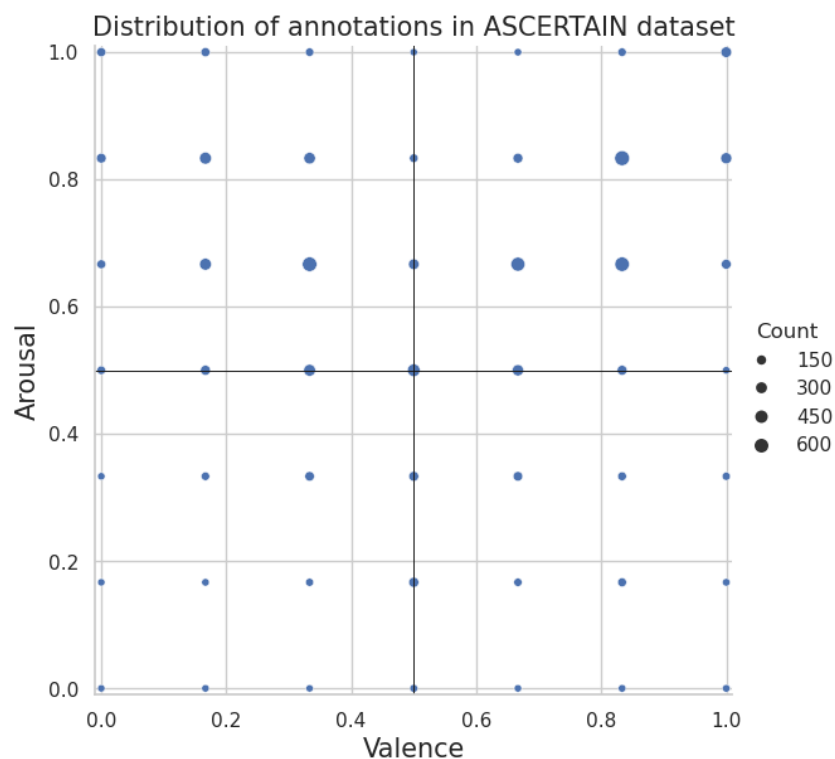


Figure 9.2: Distribution of annotations in a subset of ASCERTAIN [227] dataset used in experiments. All values were normalized relative to ranges of scales in employed questionnaires.

The original dataset consists of data from 58 university students (21 females; M age = 30). Subjects were exposed to a total of 36 movie video clips, each targeting a specific quadrant of the two-dimensional arousal-valence plane (9 videos per quadrant). Participants provided their emotions after watching stimuli, using 7-point scales to rate their perceived arousal and valence and two-point scales to

measure engagement, liking, and familiarity. During all inductions, participants had their PNS activity measured using electrocardiography, electrodermal activity, electroencephalography (only frontal pole), and facial expression features. All PNS measures are collected using simplified interfaces registering signals using only a couple of electrodes (one or two, depending on the device). Although used devices are not suitable for everyday life studies, they can be seen as a step towards real-life emotion research. The authors also provide personality scores for the Big 5 Personality model for each participant and a table indicating the data quality of each collected sample in the one to six range, with one marking excellent and six marking very poor quality.

When exploring the dataset, we noticed that some of the participant's registered signals were of very low quality. We decided to exclude recordings with a rating of three or more (data of seemingly low quality). Additionally, we excluded participants who had less than ten such samples, resulting in 10 subjects being excluded entirely. The resulting distribution of emotion annotations is presented in Fig. 9.2.

CASE

We have already described the CASE [212] dataset earlier in Sec. 8.2.1, as we used it for conducting the Emotion Physiology and Experience Collaboration (EPiC) challenge. Unlike the earlier experiments, in this research, we have not utilized the full scope of continuous annotations but divided each recording into non-overlapping windows centered around the respective annotations. The resulting distribution of emotion annotations is presented in Fig. 9.3.

DREAMER

Authors of DREAMER [228] dataset collected data from 25 subjects (11 females; M age = 26.6 years, SD = 2.7), who were exposed to a total of 18 emotionally stimulating videos, targeting states of amusement, excitement, happiness, calmness, anger, disgust, fear, sadness, and surprise (two videos for each emotion). Participants provided ratings of their emotional experience in terms of arousal, valence, and dominance (each scored from one to five) after each stimulus. Due to technical issues, two females were not included in the dataset. Thus, it consists of data from 23 subjects. Participants' PNS activity was measured during all inductions using two physiological signals, namely electrocardiography (one-channel, 256Hz) and electroencephalography (14-channel, 128Hz). Both devices can be viewed as

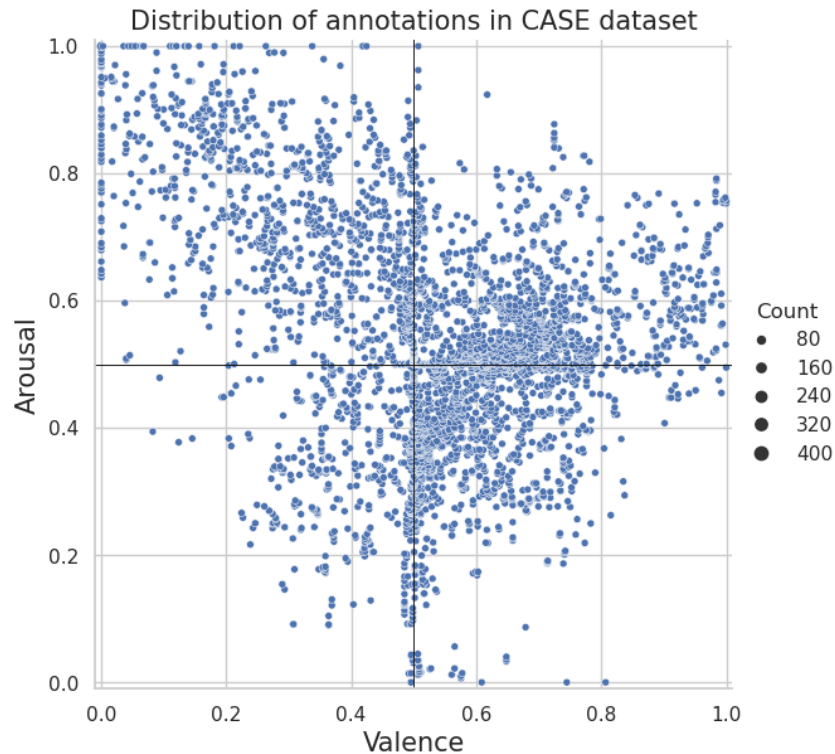


Figure 9.3: Distribution of annotations in a subset of CASE [212] dataset used in experiments. All values were normalized relative to range of scales in employed questionnaires.

relatively simple when compared with medical-grade interfaces. We present the resulting distribution of emotion annotations in Fig. 9.4.

9.1.2 Dataset preprocessing

Utilized datasets differed from one another; therefore, we had to preprocess them into one common form. First, to increase the number of samples, all recordings of emotion inductions were divided into 10-second long windows with an overlap of five seconds. We created five such windows from one induction for AMIGOS, ASCERTAIN, and DREAMER, which all have a single emotion score annotated at the end of each recording. In the CASE dataset, we utilized continuous annotations and created 22 samples per recording, restricting their total length to match that of the shortest stimulus. In all datasets for each induction, we created samples starting at their end and moving in time towards the beginning. Windowed ECG recordings were next downsampled to 100Hz to match the frequency expected by the utilized model (Sec. 9.1.4) and filtered using a high-pass Butterworth filter at 0.5Hz (order = 5) followed by powerline filtering at 50Hz. Next, samples were further processed

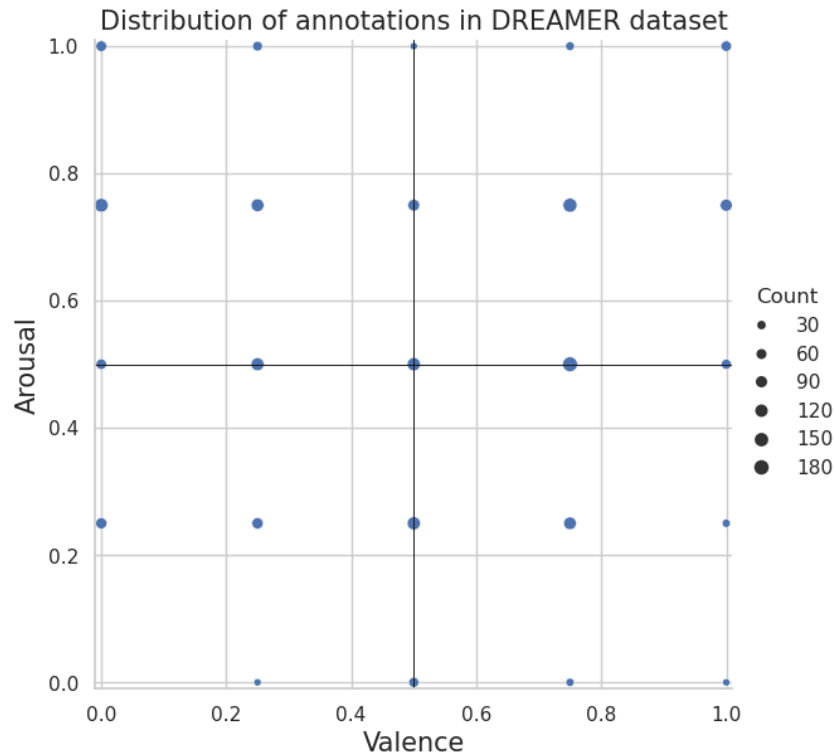


Figure 9.4: Distribution of annotations in a subset of DREAMER [228] dataset used in experiments. All values were normalized relative to range of scales in employed questionnaires.

using studied methods.

9.1.3 Studied processing methods

In this research, we investigated different approaches to personalized data processing. The approaches below differ in the produced population-subject relationships, thus affecting the quality of the obtained reasoning models. By facilitating the within-subject differences between samples collected during different emotional states, we expected models to learn this information more easily and thus achieve higher recognition accuracy. Moreover, since we preserved the original shape of signals and absolute scales of processed inputs and outputs, we expected group-personalized models to perform well by learning subject-specific differences between emotional states and population-wise physiological patterns.

All of the research methods are based on standard scoring (also known as standardization or z-scoring) of data to explore which approach to standardization yields the best results in terms of regression and classification quality. Each method was tested in a setup where test subject data were available both in training and test datasets

without knowledge leaks (multiple data windows from one session had to be all included in the training or test dataset). This allowed employing machine learning models to capture each subject's characteristics during the training phase and those relationships for reasoning during testing. In our experiments, we considered the following methods:

Method 1 Simple processing where normalization was done according to the range of recorded values. It was done by first subtracting the range's middle value and later dividing by the range (*maximum – minimum*). We also call it not standardized processing, and abbreviate it as NS.

Method 2 Standardization done per data sample, i.e., z-scoring each data sample using its own statistics. Conceptually, we treat each data sample individually in this method, so models cannot utilize first-order statistics for learning. We also call it within-sample processing and abbreviate it as WS. Applicable only to physiological signals.

Method 3 Subject-specific standardization, where each subject's data were z-scored using statistical measures calculated only from data of the given subject. Conceptually, this method aims to make within-subject physiology more regular while ignoring between-subject relationships. We also call it within-person processing and abbreviate it as WP.

The above methods were applied to both physiological data (input) and emotional experience annotations (target). In the case of target personalization, after applying person-dependent standardization and per-sample standardization, we scale results to the 0-1 range to improve training stability and to ensure that scales are consistent and comparable. When describing the results, we denote applied processing methods as P - *physiology processing method* & A - *annotation processing method*, e.g., P-NS&A-WP denotes non-standardized physiology and annotations standardized in a within-person manner.

When planning the experiments, we also considered other approaches, including standardization respective to population-wise statistics and two-stage methods, e.g., population-wise standardization together with additional subject-wise processing. However, when using standardization (or any other transformation based on shifting and scaling), each new transformation applied to the already transformed dataset (or

to its subset) overrides the effects of previously applied transformations (mathematical reasoning supporting this conclusion can be found in Appendix Sec. E.1). This cancelation does not apply to transformations applied to smaller subsets first and later to the whole dataset, which we utilize when scaling annotations standardized in a person-dependent fashion to the 0-1 range.

The above methods were tested both by themselves (e.g., physiology processed using min-max approach with annotations processed in a subject-specific way) and together with other methods from the list (e.g., both physiology and annotations processed using subject-specific approach) on emotion classification and regression tasks. For the classification task, after processing, we divided data into binary high emotion (1) and low emotion (0) states around the middle value from the range (0.5). For regression, after processing with the considered method, emotion annotations were scaled to the 0-1 range and used as a ground truth.

9.1.4 Emotion modeling

To examine the above data processing methods for emotion modeling, we employed the *WildECG* model that we proposed in one of our articles [29]. It is based on an S4 state-space model [230] pre-trained on the *Tracking Individual Performance with Sensors* (TILES) [231] dataset – one of the most extensive publicly available outside-the-laboratory biosignal datasets.

To learn to extract knowledge, during pretraining the model was tasked with recognizing transformations applied to ECG signal samples in a multi-label classification setting. The set of possible transformations consisted of (1) masking, (2) cropping, (3) noise addition (white noise or random wander), (4) signal permutation, (5) time warping (stretching or squeezing), (6) magnitude scaling, (7) inverting signal along the temporal axis, and (8) reversing the signal (multiplication by -1). During pretraining, each augmentation was applied with a predefined probability, and at most, four transforms were applied simultaneously. More details on transformations and training procedures can be found in the published article [29].

9.2 Results

In this section, we compare different processing strategies in group- and individual-model designs. Due to the relatively low number of samples in datasets, all models were trained and tested using 5-fold cross-validation. Initially, we conducted experiments using two training schemas, i.e., training the whole model (trained all / fully trained) or only the downstream task head (trained head) with frozen pre-trained

weights. To simplify analyses, we decided to compare processing methods using one training schema (trained all) chosen based on obtained results from experiments conducted with only basic preprocessing applied (NS).

All of the provided accuracy measurements for models are accompanied by baseline results, namely majority label or average value (respectively for classification and regression) from test sets (fold) to check if models can predict emotional states better than naive predictors. For comparisons of modeling approaches in a group-wise scenario, we also utilized non-initialized models trained solely in a supervised manner, as in other experiments, personal datasets were too small in size to train deep learning models solely on them.

In this section, we also describe our attempt at comparing different experimental designs, i.e., creating general, group-personalized, and subject-specific models and their impact on obtained model accuracy. For better comparability, we analyzed results in a subject-wise manner by first averaging results from five folds within a person and comparing such population-wise distributions between experiment setups.

The obtained results were tested for statistical significance. Comparisons with baselines were performed using Wilcoxon’s signed-rank test [232]². Comparisons between different studied methods were performed using Friedman’s test for χ^2 and F distributions [233], followed by Conover’s post-hoc test [234, 235]³. To control family-wise error rate, all of obtained p-values were corrected using Holm–Bonferroni procedure [209]. While Wilcoxon’s tests were performed on raw values of metrics (only computed from obtained predictions), for comparisons between processing methods, we had to account for differences in annotations caused by different annotation processing procedures (as shown in Fig. 9.5). Since those differences are well visible in baseline scores, we performed statistical tests on distributions of *gain* values. For M components (e.g., number of folds over which we compute results), vectors of baseline scores $B = (B_1, \dots, B_M)$, and prediction scores $P = (P_1, \dots, P_M)$, vector of gain values G can be expressed as:

$$G = \left(\frac{P_1 - B_1}{B_1}, \dots, \frac{P_M - B_M}{B_M} \right)$$

²Computed using rstatix R package, v. 0.7.2.

³Computed using PMCMRplus R package, v. 1.9.10.

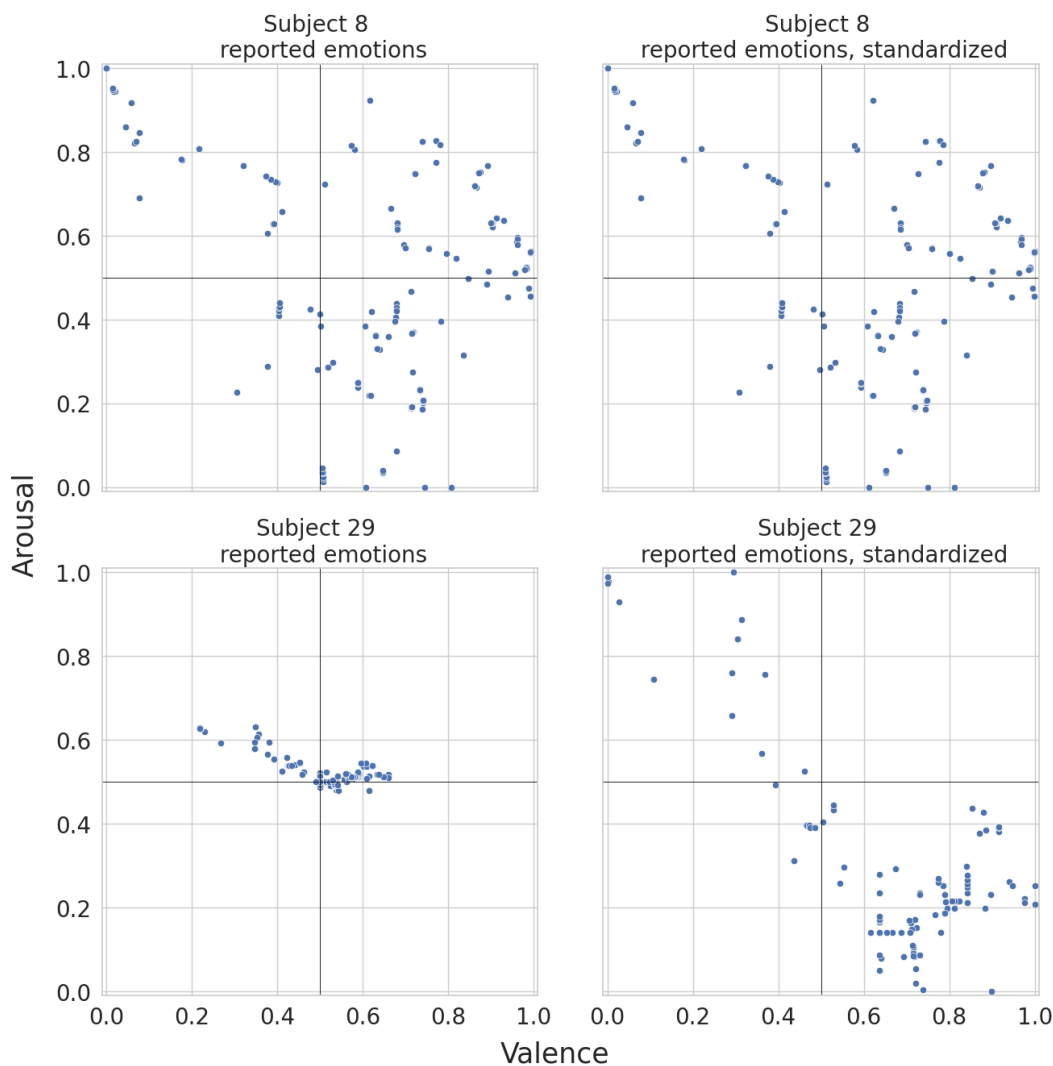


Figure 9.5: Self-reported arousal and valence of selected subjects from CASE [212] dataset before and after adjusting.

9.2.1 Processing strategies for group models

When training in group setup for classification task, adjusting all weights of a model in most cases resulted in higher metric values than training only classification head with pre-trained weights frozen (Fig. 9.6, Tab. E.1). Also, all fully-trained classification models beat baseline models in terms of achieved recognition accuracy (macro averaged F1, precision, and recall measures). On the other hand, in regression task models not only struggled to achieve higher (or even the same) prediction accuracy than average baselines (RMSE, MAE), but it is hard to determine which training strategy was better (Fig. 9.7, Tab. E.2). However, in terms of CCC measure, fully-trained models performed the best, although we cannot tell if better than base-

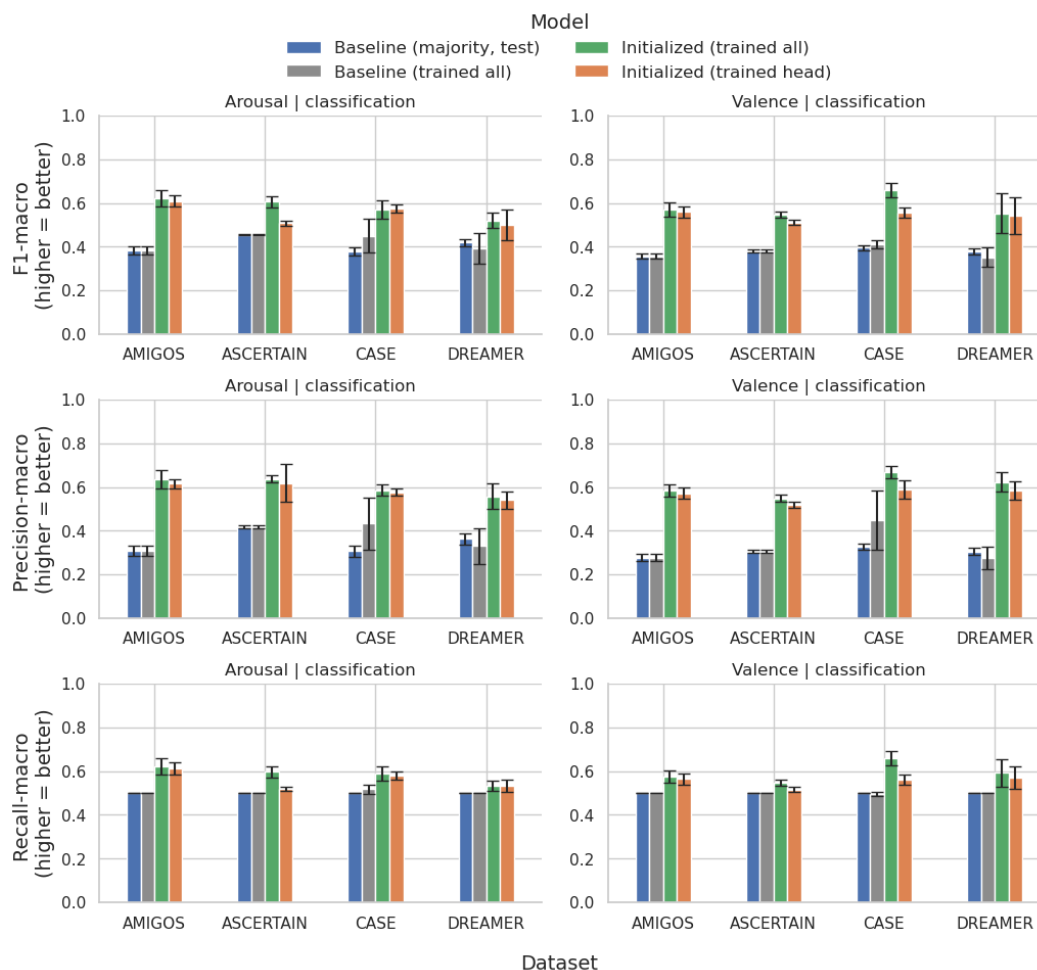


Figure 9.6: Comparison of model training approaches with baselines on classification tasks in group setup. Initialized - models initialized with pre-trained weights; trained all - training all weights of a model; trained head - training only downstream task head, while leaving rest of weights frozen. For specific values see Tab. E.1.

lines (CCC equals zero for uniform predictions). For further analyses we selected training strategy where all weights were trained, as it (1) proved to be the best in classification setting, and (2) achieved highest CCC scores in all cases.

AMIGOS dataset

Results for the AMIGOS dataset show that predictions from models were, in the majority of cases, different from respective baselines (Fig. 9.8). While in classification tasks, such result comes from outperforming baseline predictors, in regression tasks, no model managed to achieve lower error than baseline values, with only valence predictors for within-person standardized annotations producing results that

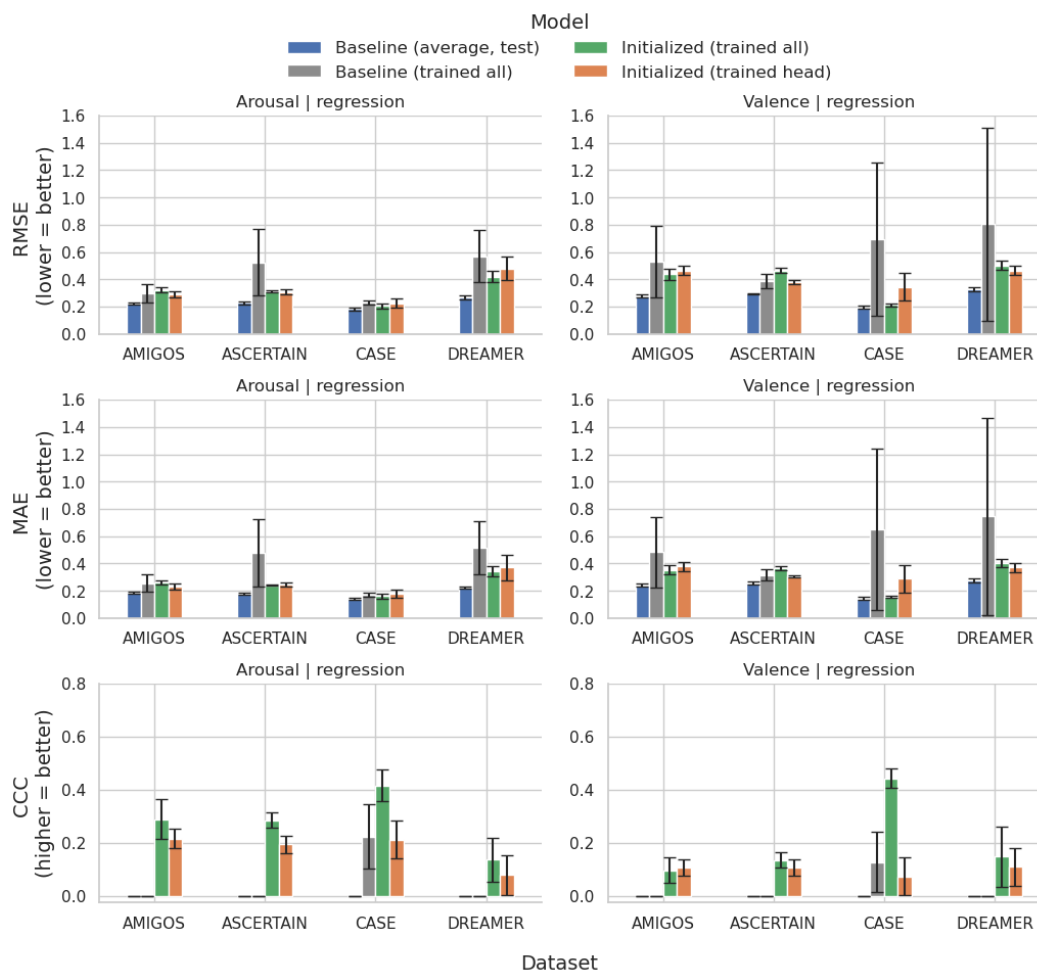


Figure 9.7: Comparison of model training approaches with baselines on regression tasks in group setup. Initialized - models initialized with pre-trained weights; trained all - training all weights of a model; trained head - training only downstream task head, while leaving rest of weights frozen. For specific values see Tab. E.2.

are not significantly worse than baseline.

Although some differences in achieved values can be observed between all processing methods, Friedman's test showed that they are statistically significant only in arousal classification task (Tab. E.3), which was confirmed by further post-hoc pairwise comparisons. We observed slightly higher values of F1 measure and gain for models trained on non-standardized annotations when compared with corresponding models trained on standardized annotations and found significant differences between 9 out of 15 pairs of tested processing methods (Tab. E.6). These results suggest that the main factor influencing differences between models is the annotation processing method, with differences between models utilizing the same processing

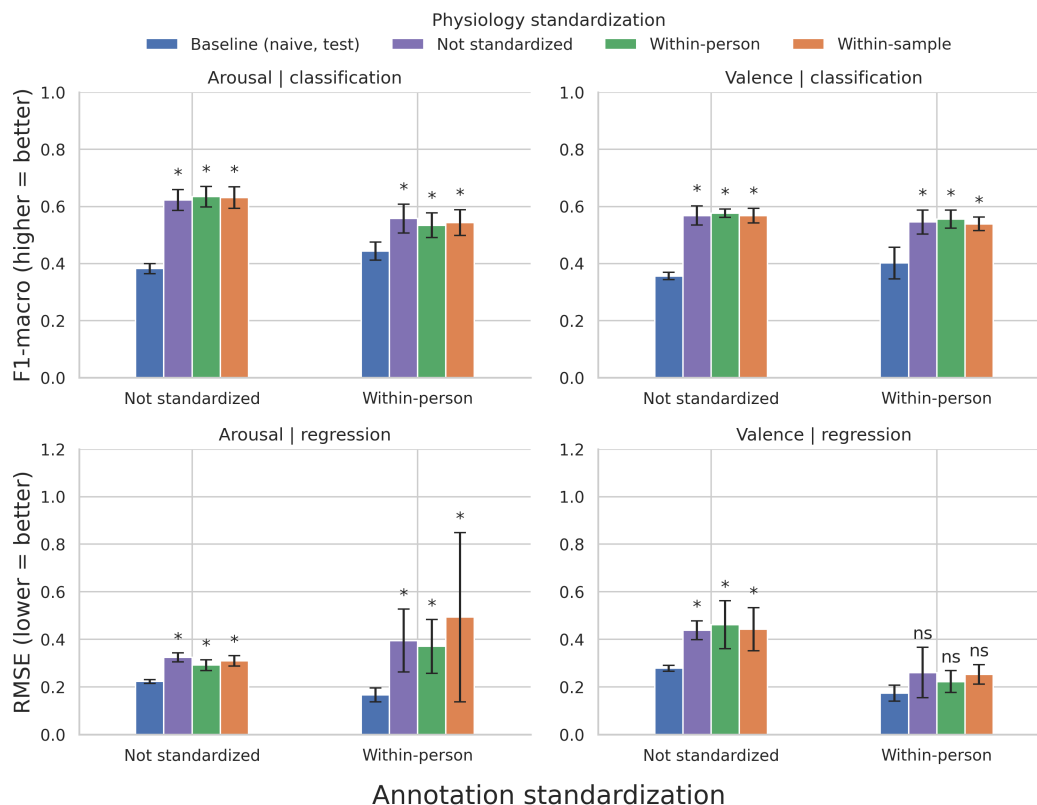


Figure 9.8: Comparison of processing methods on AMIGOS dataset in group training setup. Baseline - majority class (for classification) or average value (for regression) from test set. Annotations above error bars denote results of the Wilcoxon's test for statistical significance of differences between methods and baselines: * - $p < 0.05$, n.s. - not significant. For specific values see Tabs E.4 and E.5.

being always statistically insignificant and between models utilizing different processing being always statistically significant.

ASCERTAIN dataset

Results for the ASCERTAIN dataset show that predictions from models were always different from respective baselines (Fig. 9.9). Similarly to the AMIGOS dataset, in classification tasks, such results come from outperforming baseline predictors. In contrast, in regression tasks, no model managed to outperform baselines or achieve the same error rates.

Friedman's test results showed that only for arousal regression, all processing methods yielded results with no statistically significant differences between them (Tab. E.3). In a classification of emotion states, we observed universally higher F1 macro and gain scores for the prediction of non-standardized emotion annotations.

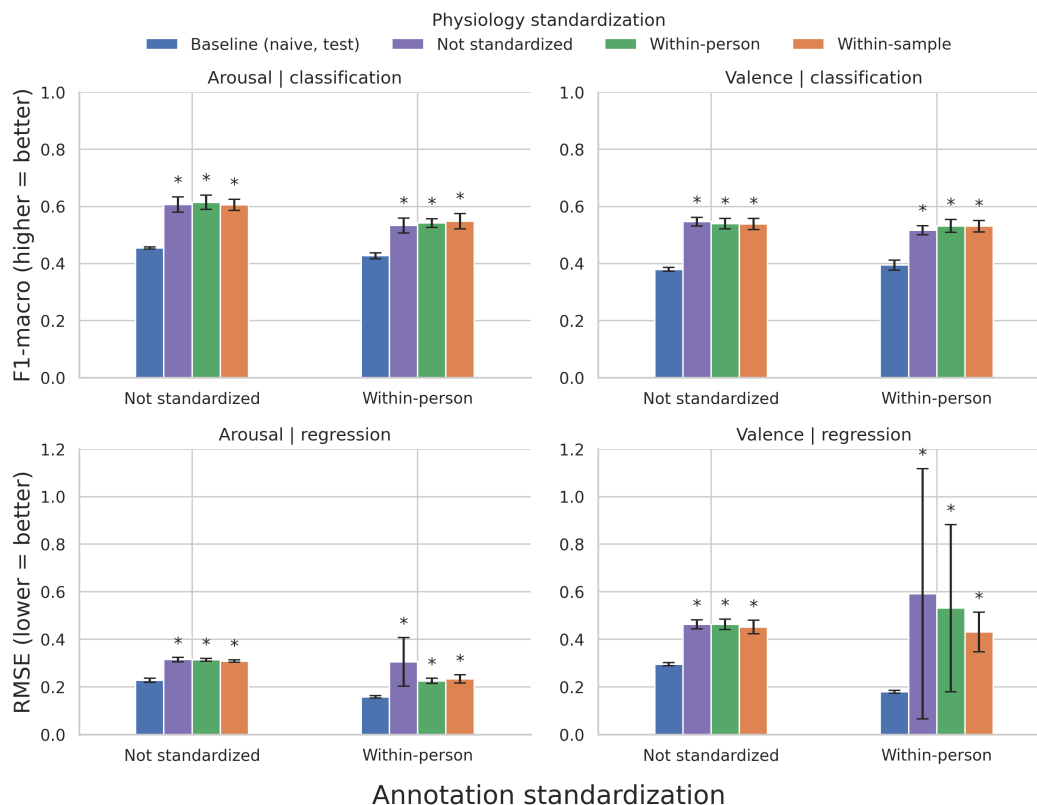


Figure 9.9: Comparison of processing methods on ASCERTAIN dataset in group training setup. Baseline - majority class (for classification) or average value (for regression) from test set. Annotations above error bars denote results of the Wilcoxon's test for statistical significance of differences between methods and baselines: * - $p < 0.05$. For specific values see Tabs E.7 and E.8.

For each of the classification tasks, we observed two differing pairs of methods. In neither of those tasks were we able to identify a common factor impacting results (Tab. E.9, E.10).

Regression of valence achieved lower (better) error rates when predicting non-standardized values for two out of three physiology processing methods ($P-NS$ and $P-WP$). Also, it always resulted in lower gain scores due to a much higher baseline error value. Also, for valence regression, we identified three pairs of processing methods that showed significant differences in scores. Differently from results for the AMIGOS dataset, the only pattern that we observed was that combining non-standardized physiology with within-person standardized annotations led to the highest, although never statistically significant, average error rates in both regression tasks (Tab. E.11).

CASE dataset

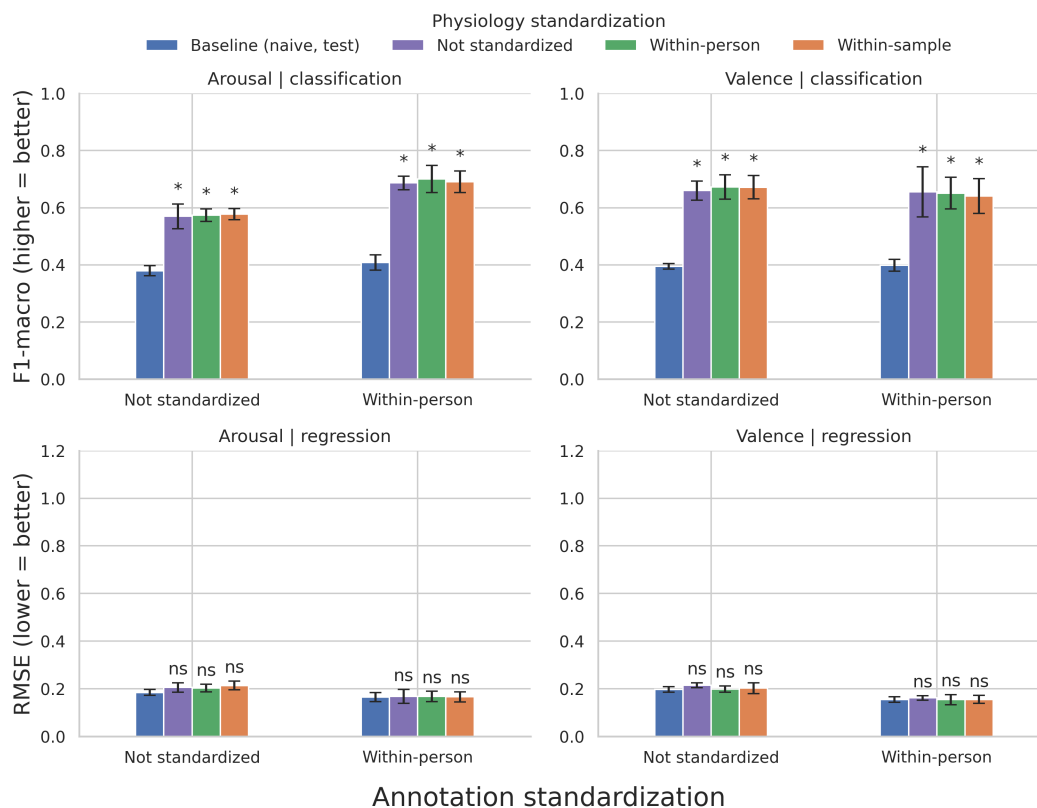


Figure 9.10: Comparison of processing methods on CASE dataset in group training setup. Baseline - majority class (for classification) or average value (for regression) from test set. Annotations above error bars denote results of the Wilcoxon's test for statistical significance of differences between methods and baselines: * - $p < 0.05$, n.s. - not significant. For specific values see Tabs E.12 and E.13.

For the CASE dataset, we again observed models performing much better than the baseline in both classification tasks (Fig. 9.10). However, for all regression tasks, error rates achieved by models were statistically no different from baselines (Fig. 9.10), which led to lower relative error rates when compared with other datasets.

Results of Friedman's test showed that only for arousal regression can we expect some differences between processing methods (Tab. E.3). Indeed, pairwise post-hoc comparisons showed that their results were significantly different for seven pairs of processing methods' combinations (Tab. E.11). One source of differences can be found in annotation processing methods, as models using the same annotation processing never showed statistically significant differences. We can also observe that differences between predictions are never caused only by the assumed physiology processing method, as for all models utilizing the same annotation processing

method, differences in predicted values were insignificant. In this task, raw error rates were always lower for within-person standardized annotations.

DREAMER dataset

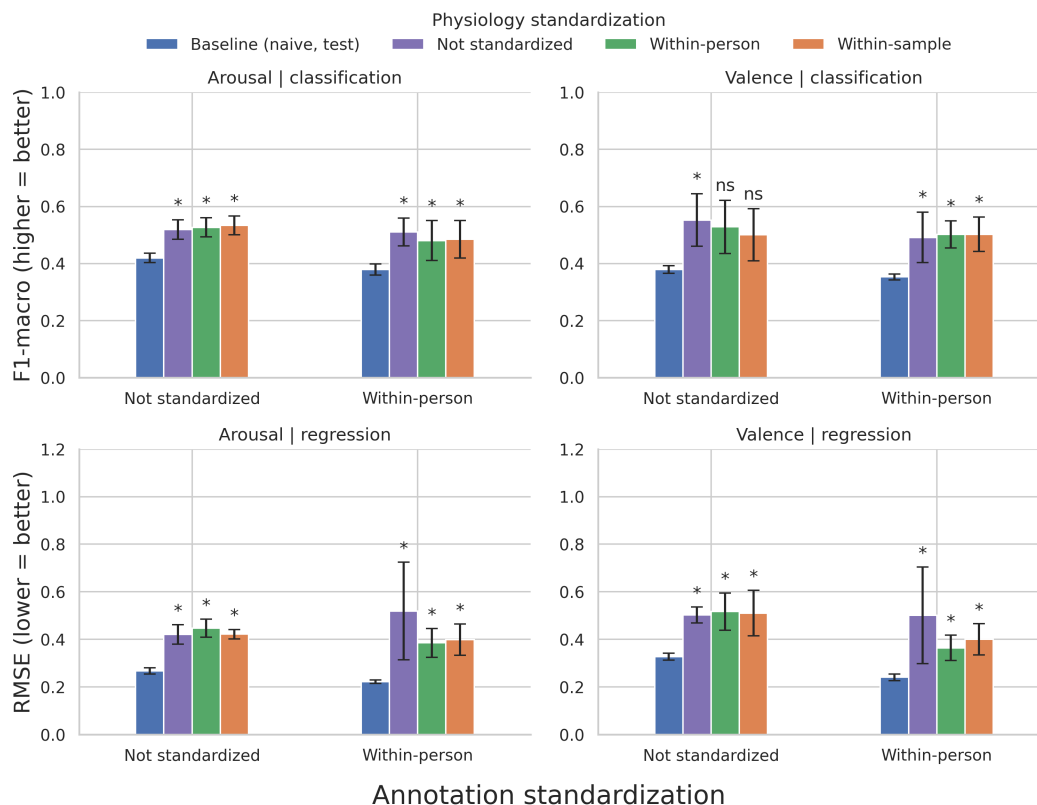


Figure 9.11: Comparison of processing methods on DREAMER dataset in group training setup. Baseline - majority class (for classification) or average value (for regression) from test set. Annotations above error bars denote results of the Wilcoxon's test for statistical significance of differences between methods and baselines: * - $p < 0.05$, n.s. - not significant. For specific values see Tabs E.15 and E.16.

For the DREAMER dataset, differently from results on other datasets, not all classification models produced results statistically different from baseline, i.e., in valence classification for non-processed self-assessment scores, within-person, and within-sample physiology processing resulted in models showing no significant differences from the majority class model. For regression tasks, all models' error rates were higher than those of baseline predictors (Fig. 9.11). However, for no task on this dataset did Friedman's test show significant differences between processing methods (Tab. E.3).

9.2.2 Processing strategies for subject-specific models

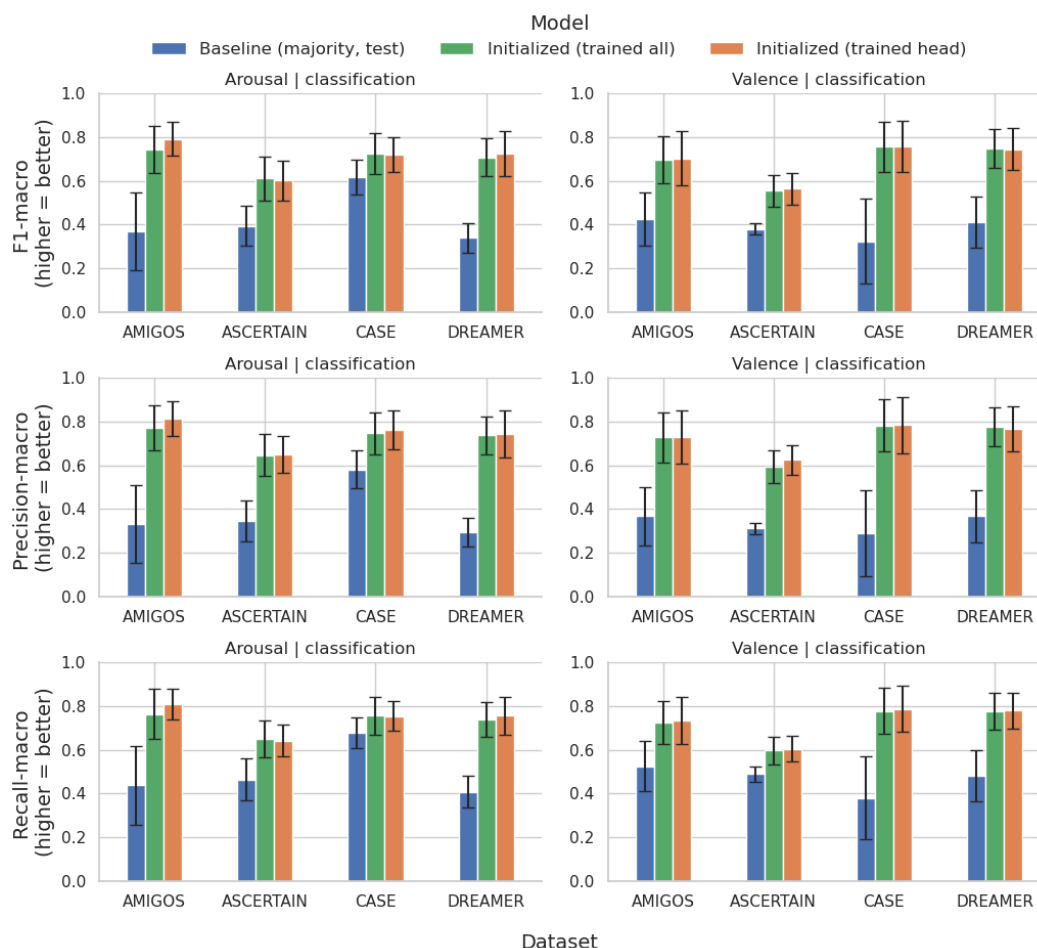


Figure 9.12: Comparison of model training approaches with baselines on classification tasks in subject-specific setup. Initialized - models initialized with pre-trained weights; trained all - training all weights of a model; trained head - training only downstream task head, while leaving rest of weights frozen. For specific values see Tab. E.17.

In subject-specific setup, training whole models usually resulted in very similar or the same results as training only classification head with pre-trained weights frozen, with the latter ones more often achieving better performance (Fig. 9.12, Tab. E.17). In the regression task, same as in the group modeling setup, no model achieved better performance than the baseline in terms of RMSE and MAE measures. However, fully-trained models almost always achieved smaller errors and higher CCC than models with frozen pre-trained weights (Fig. 9.13, Tab. E.18). Therefore, for further analyses, we selected a training strategy where all weights were trained.

Results for comparisons of trained models with baselines show that subject-specific

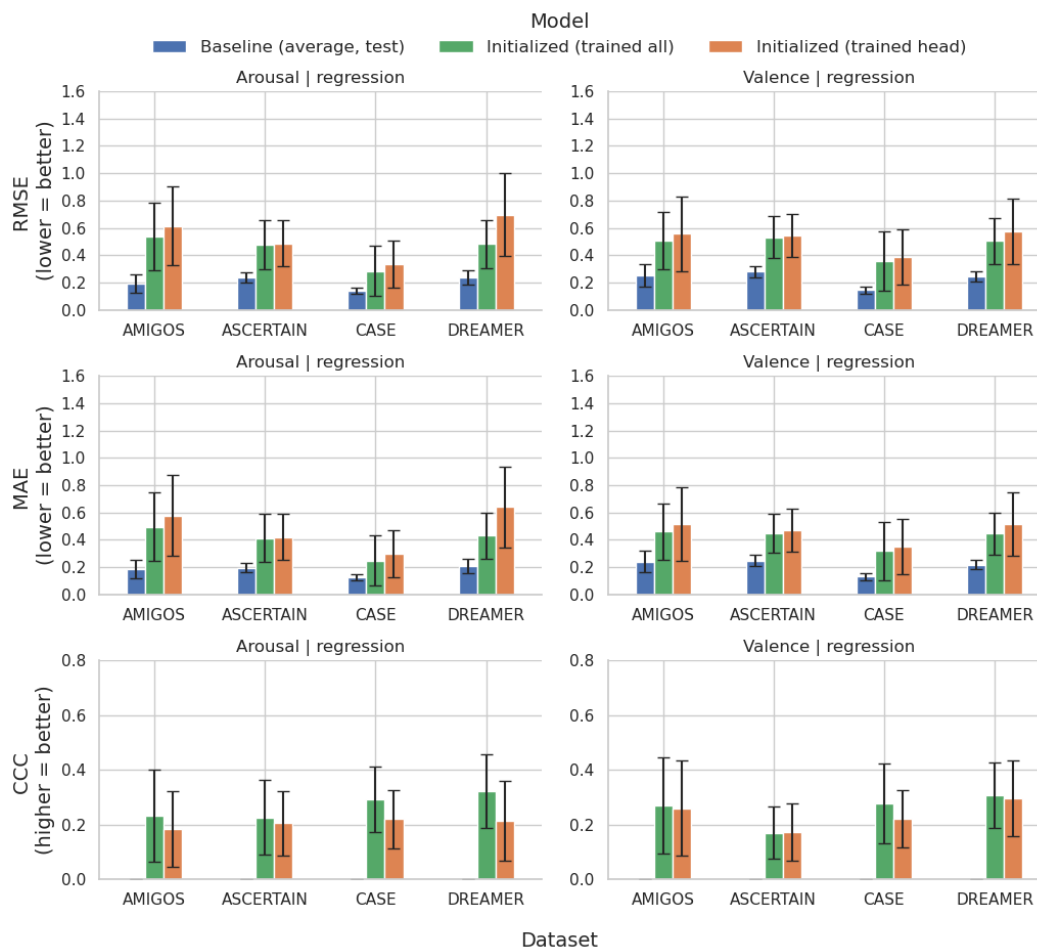


Figure 9.13: Comparison of model training approaches with baselines on regression tasks in subject-specific setup. Initialized - models initialized with pre-trained weights; trained all - training all weights of a model; trained head - training only downstream task head, while leaving rest of weights frozen. For specific values see Tab. E.18.

models consistently achieved significantly higher classification accuracy (in terms of F1 score) and significantly higher error in regression tasks (in terms of RMSE) than corresponding baselines. From the results of Friedman's test, we expected significant differences between at least two utilized processing methods in 9 out of 16 studied cases (Tab. E.19).

AMIGOS dataset

For the AMIGOS dataset, Friedman's test showed expected significant differences between processing methods only for regression tasks (Tab. E.19). We noticed that regression of non-standardized emotion levels always achieved lower RMSE than re-

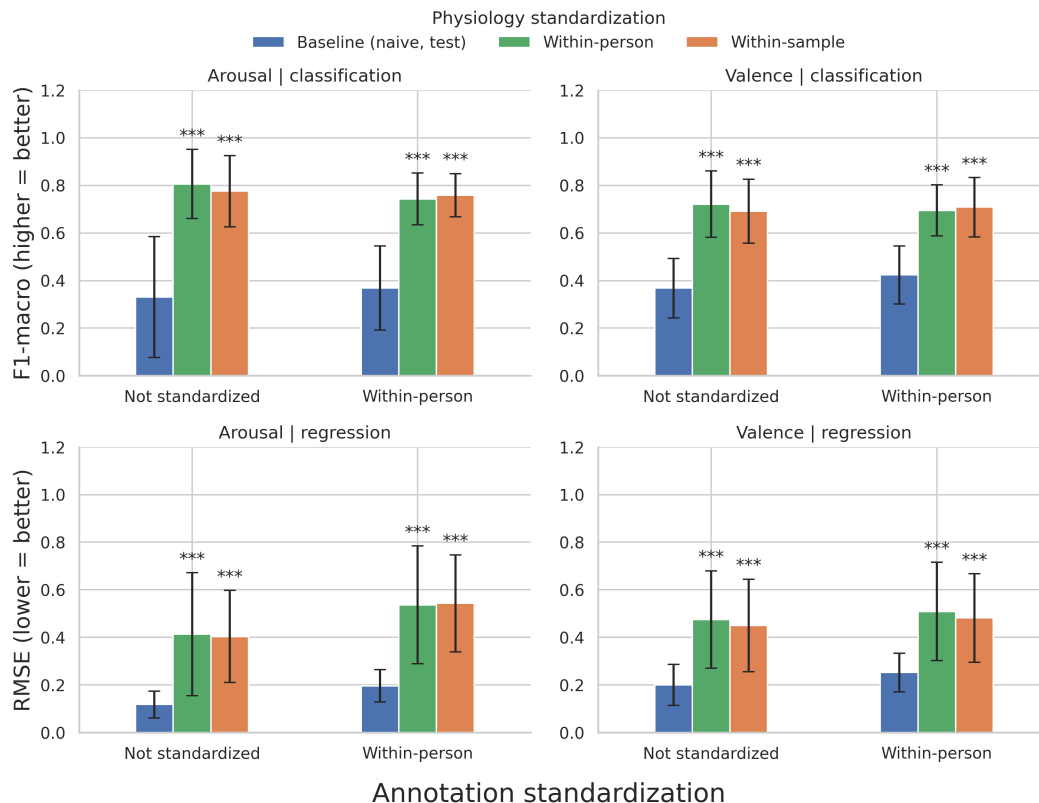


Figure 9.14: Comparison of processing methods on AMIGOS dataset in subject-specific training setup. Baseline - majority class (for classification) or average value (for regression) from the test set (computed subject-wise). Annotations above error bars denote results of the Wilcoxon's test for statistical significance of differences between methods and baselines: *** - $p < 0.001$. For specific values see Tabs E.20 and E.21.

gression of within-person standardized values while achieving higher (worse) gains due to lower baseline error rates (Fig. 9.14). In Conover's post-hoc test for arousal recognition, we can notice that most of the applied methods showed significant differences, which were driven mainly by different approaches to annotation processing (Tab. E.22). Significant differences were found only for one of the tested pairs of processing approaches for the valence task, so we could not draw any conclusions (Tab. E.23).

ASCERTAIN dataset

Friedman's test applied to results from the ASCERTAIN dataset showed that we should expect significant differences between methods in all tasks but valence regression (Tab. E.19). Emotion classification models consistently achieved higher

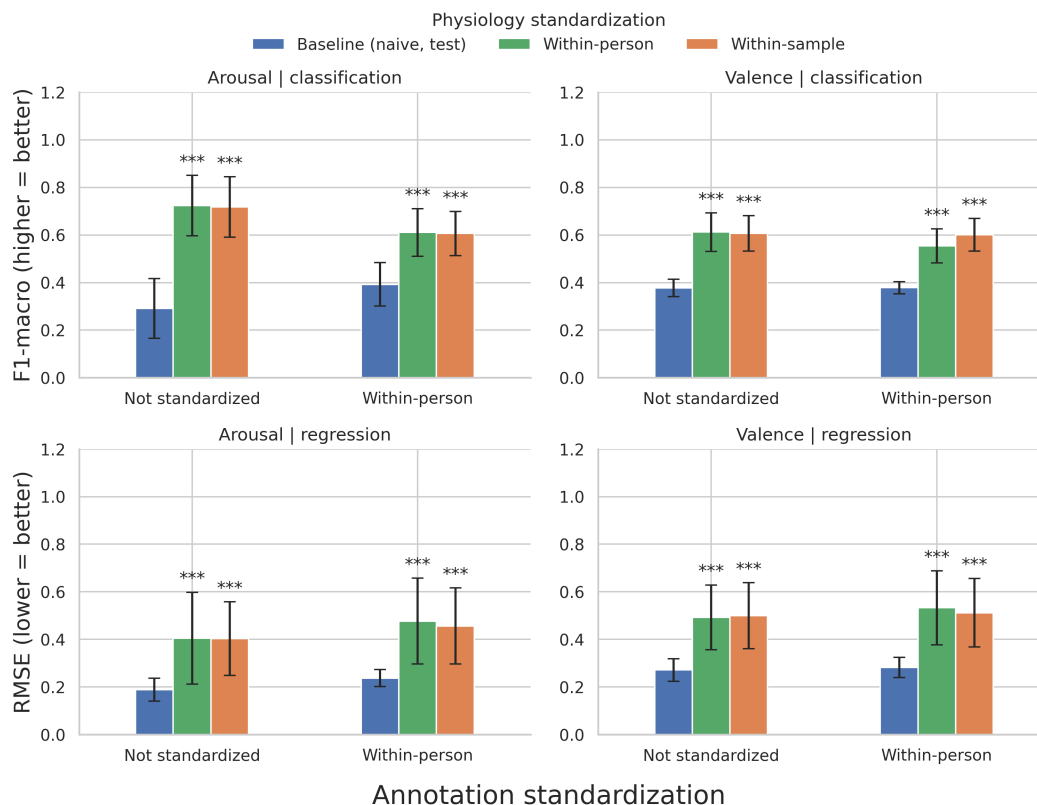


Figure 9.15: Comparison of processing methods on ASCERTAIN dataset in subject-specific training setup. Baseline - majority class (for classification) or average value (for regression) from the test set (computed subject-wise). Annotations above error bars denote results of the Wilcoxon's test for statistical significance of differences between methods and baselines: *** - $p < 0.001$. For specific values see Tabs E.24 and E.25.

F1-macro and gain values when predicting non-standardized emotions (Fig. 9.15). Similarly, arousal regression models achieved lower error rates for non-standardized annotations. However, they also achieved higher gains due to lower baseline RMSE.

Contrary to what we expected from Friedman's test, Conover's post-hoc for arousal regression showed no statistically significant differences between processing methods (Tab. E.28). For arousal classification, results show similar behavior as in arousal regression on the AMIGOS dataset, with no significant differences between models trained on annotations processed in the same manner (Tab. E.26). However, we also observed that in this task, differences between $P-WP\&A-WP$ and $P-WS\&A-NS$ were of no statistical significance. For valence classification, we noticed that statistically significant differences were present only between $P-WP\&A-WP$ and other methods, with said method achieving the lowest F1-macro and gain (Tab. E.27).

CASE dataset

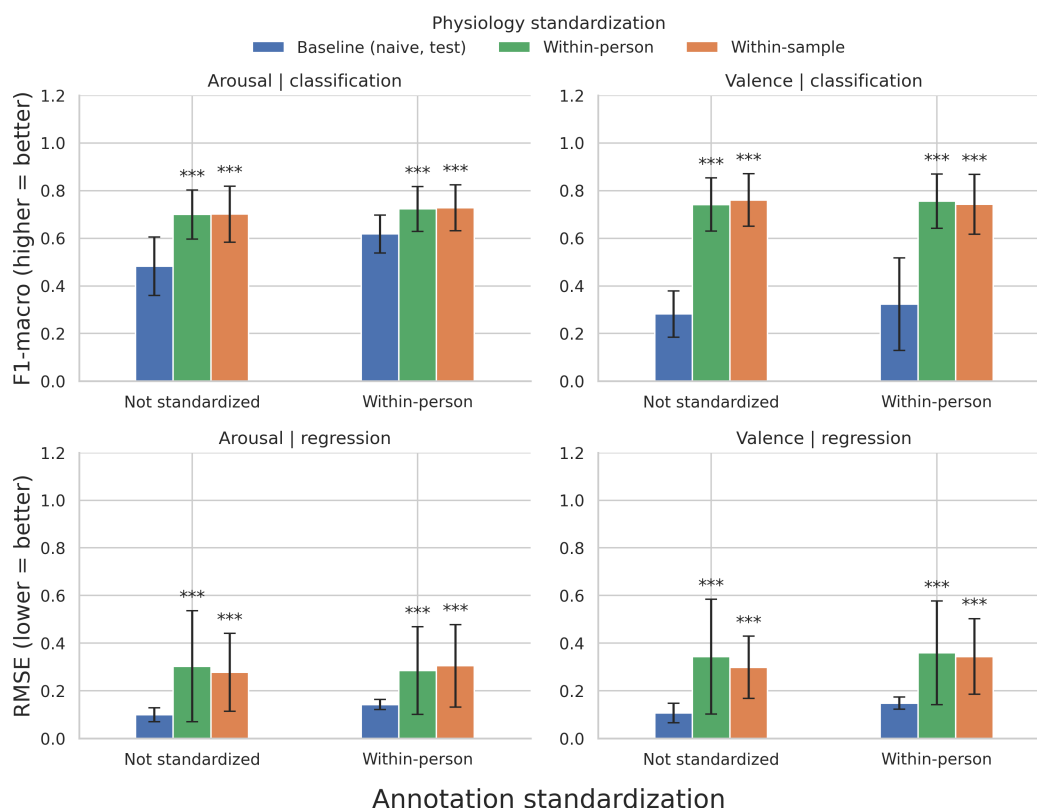


Figure 9.16: Comparison of processing methods on CASE dataset in subject-specific training setup. Baseline - majority class (for classification) or average value (for regression) from the test set (computed subject-wise). Annotations above error bars denote results of the Wilcoxon's test for statistical significance of differences between methods and baselines: *** - $p < 0.001$. For specific values see Tabs E.29 and E.30.

Results of Friedman's test for the CASE dataset show expected significant differences only for arousal recognition, both as classification and regression tasks. Models for classification of within-person standardized arousal achieved higher F1-macro than their non-standardized counterparts (Fig. 9.15). In arousal regression, the P-WP&A-WP model showed similar results, achieving lower RMSE than the P-WP&A-NS model. We also noticed that predictors for non-standardized arousal always achieved higher gain values in both classification and regression. Also, Conover's test results showed that only methods differing in annotation processing produced results that were significantly different (Tabs. E.31 and E.32).

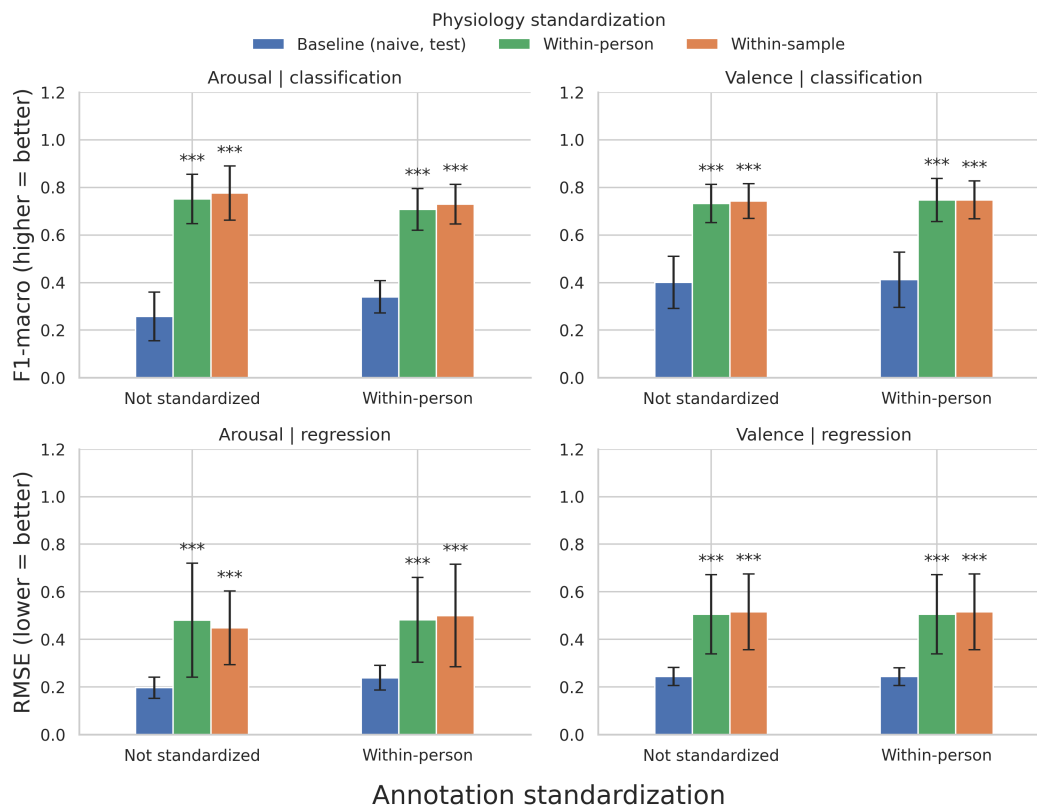


Figure 9.17: Comparison of processing methods on DREAMER dataset in subject-specific training setup. Baseline - majority class (for classification) or average value (for regression) from the test set (computed subject-wise). Annotations above error bars denote results of the Wilcoxon's test for statistical significance of differences between methods and baselines: *** - $p < 0.001$. For specific values see Tabs E.33 and E.34.

DREAMER dataset

Similarly, as for the CASE dataset, Friedman's test shows expected significant differences only for arousal recognition in classification and regression tasks. However, for the DREAMER dataset, we noticed that predicting non-standardized emotions resulted in better values of metrics (higher for F1-macro, lower for RMSE; Fig. 9.15). In the regression task, Conover's test showed no significant differences between pairs of processing methods (Tab. E.36) and in the classification task only $P\text{-}WP\&A\text{-}WP$ and $P\text{-}WS\&A\text{-}NS$ differed on a statistically significant level.

9.2.3 Comparison of experimental designs

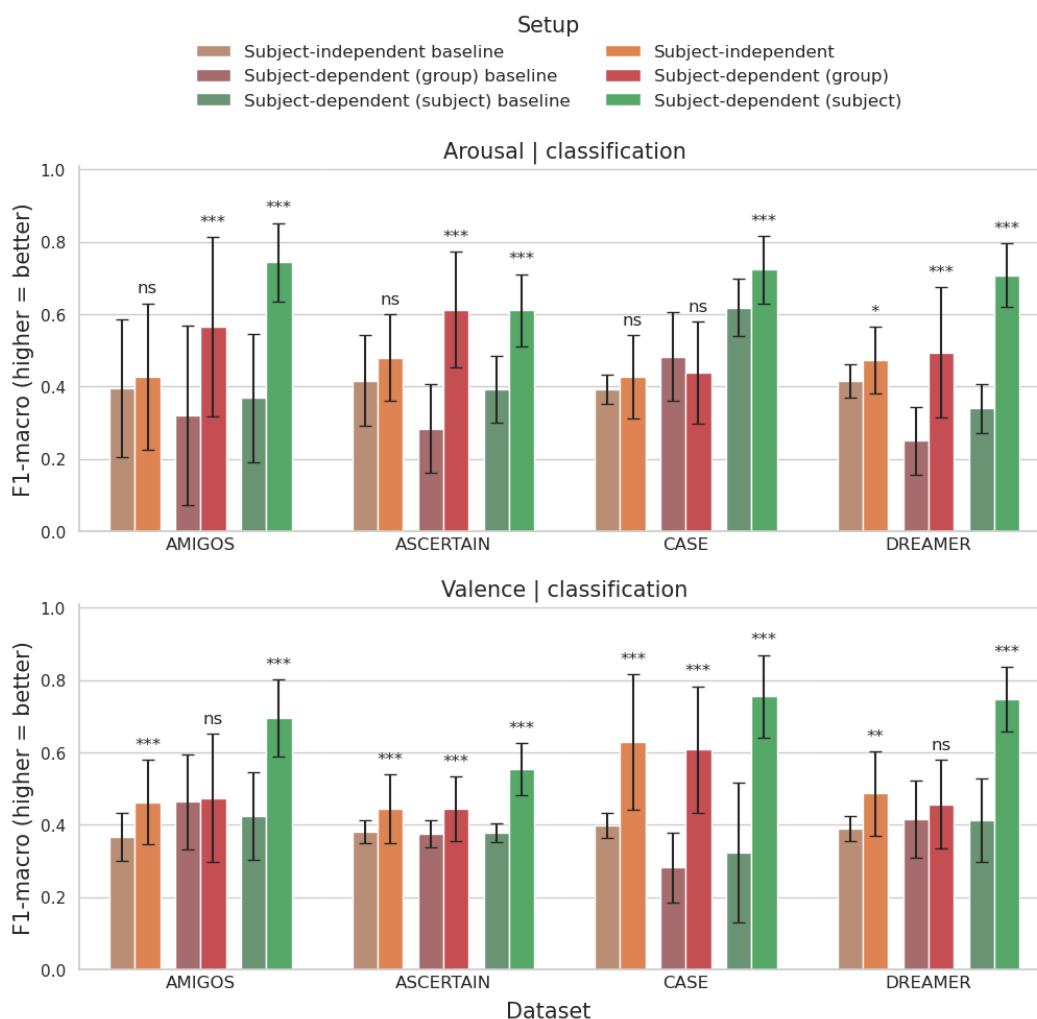


Figure 9.18: Comparison of experimental designs with baselines on classification tasks. For specific values see Tabs. E.37 and E.38.

Comparisons of experimental designs (setups) were first conducted against respective baselines, which differ from each other because training had to be conducted differently for each design. Although all average metrics were always higher for models, only subject-dependent models produced predictions that were significantly better from baselines in all classification tasks (Fig. 9.19, Tabs. E.37 and E.38). Subject-independent and group-personalized models produced predictions with no statistically significant differences from those of baseline models in three classification tasks each (subject-independent models in arousal prediction on AMIGOS, ASCERTAIN and CASE dataset; group-personalized models in arousal prediction on CASE dataset and valence prediction on AMIGOS and DREAMER datasets). In regression tasks, although some differences exist between models, in all exper-

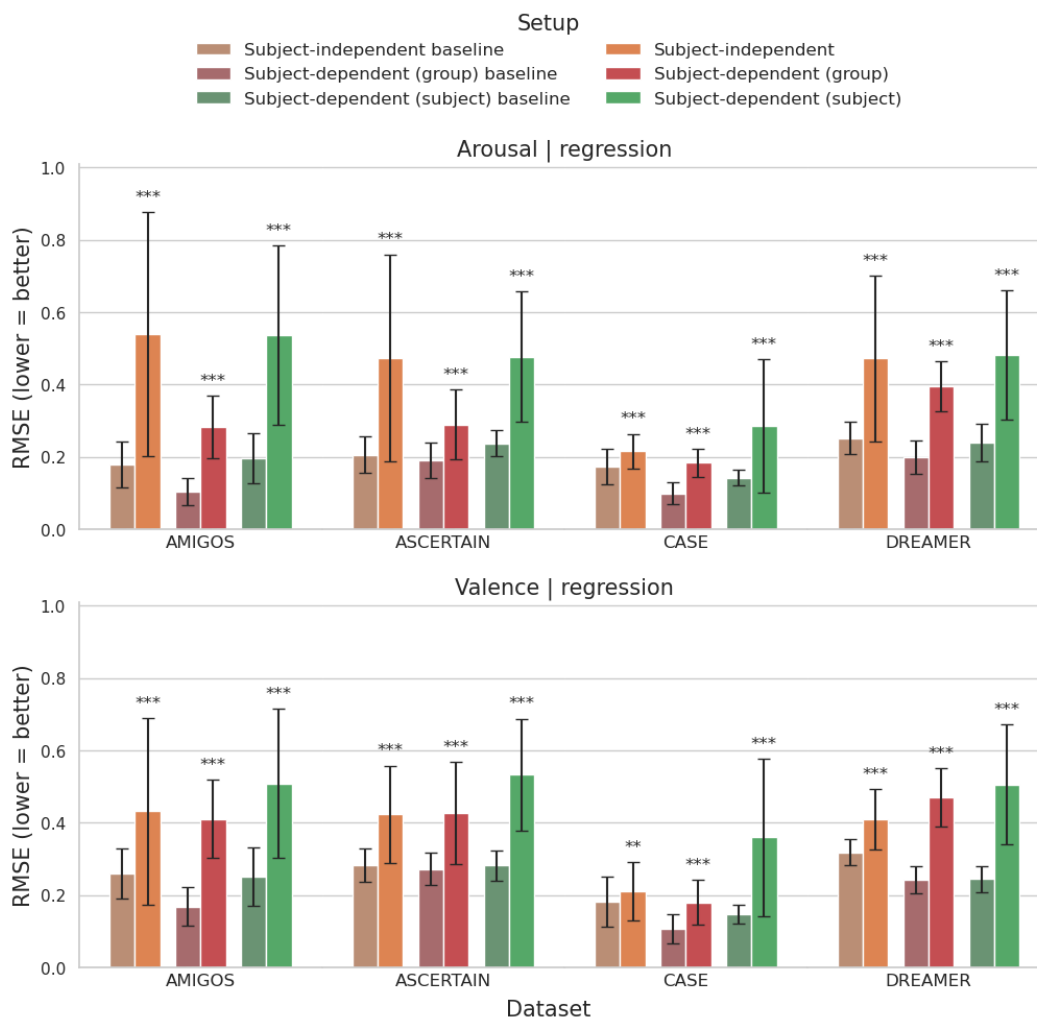


Figure 9.19: Comparison of experimental designs with baselines on regression tasks. For specific values see Tabs. E.39 and E.40.

imental designs, they achieved significantly higher error rates than corresponding baselines.

Results from Friedman's test suggested that in almost all tasks and datasets, some significant differences between experimental designs exist (except arousal regression on AMIGOS and DREAMER datasets (Tab. E.41). Further post-hoc analyses on baseline-corrected values (gains) showed that in classification tasks, personalized models (both group- and subject-personalized) always predicted arousal better than general models. In most cases it was also true for valence predictions, with no significant differences between subject-independent and group-personalized models on ASCERTAIN, CASE, and DREAMER datasets (Tabs. E.46, E.50, and E.54). Also, in arousal classification on ASCERTAIN dataset, no significant differences

were observed between predictions from subject- and group-personalized models.

Contrary to classification experiments, where personalized models almost always achieved higher prediction accuracy than general ones, in regression tasks, subject-wise models often achieved the highest (worst) average prediction error (RMSE) 9.18, and subject-independent models performed the best in valence regression on ASCERTAIN and DREAMER datasets. When accounting for baselines (gain), subject-independent models achieved the lowest prediction error in all but two tasks (arousal prediction on AMIGOS and ASCERTAIN datasets, Tabs. E.39 and E.40), albeit it was always significantly higher than baseline. Although mean scores always differed between models, Conover's post-hoc tests (Tabs. E.42-E.55) showed that subject-independent models performed similarly to personalized ones in arousal regression on ASCERTAIN dataset (subject-personalized) and valence regression on ASCERTAIN and CASE datasets (group-personalized).

9.3 Discussion

Methods in this section were researched using four laboratory datasets well-known in the literature. Unfortunately, we could not test them on our real-life dataset (Chap. 7), as due to its size and complex nature, processing and cleaning the data have not been finished at the time of writing. Also, we focused our efforts on ECG-based emotion recognition, which made our own laboratory dataset unusable due to the lack of this modality (Chap. 5).

In this chapter, we mainly focused on researching the effect of different standardization approaches on the accuracy of deep learning models for emotion recognition from ECG signals. Processing strategies were utilized to examine the effect of within-subject processing on achieved metrics. We expected that increasing between-subject differences while preserving the original signals' shapes would facilitate capturing the subject's specificity while learning population-wise patterns.

In both group- and personal-model setups, classifiers usually achieved significantly higher metric values than respective majority class baselines, but Friedman and Conover's tests often showed no statistically significant differences between classification models using different processing strategies. We observed some significant differences between arousal predictors for all datasets and between valence predictors only for the ASCERTAIN dataset. Also, these results depended on an experimental setup, with processing methods showing differences in AMIGOS and ASCERTAIN datasets for group models and ASCERTAIN, CASE, and DREAMER datasets for personal models. However, from those results, we can derive that for the classification of emotional states, annotation processing often does not matter, and in the cases where differences are present, using non-standardized annotations led to better results than within-person standardized ones. Also, we did not notice any significant differences between physiology processing methods.

Regression models showed opposite behavior to their classification counterparts, achieving usually higher error rates than average baselines. Nonetheless, we still decided to examine the results, as average predictors are usually very strong baselines in regressions tasks, as shown, for example, in Chap. 8. Same as for classification models, Friedman's and Conover's tests often showed no statistically significant differences between processing methods. We observed some significant differences between methods for arousal regression on AMIGOS, ASCERTAIN, and CASE datasets, for valence regression on AMIGOS and ASCERTAIN, and no differences between predictors for the DREAMER dataset. Also, these results depended on an

experimental setup, with processing methods showing differences in ASCERTAIN and CASE datasets for group models and AMIGOS and CASE datasets for personal models. Again, we noticed that statistically significant differences originated mainly from annotation processing methods. Group-personalized models trained on within-person standardized annotations showed lower error rates and gains for arousal on the CASE dataset and, in one case, for valence prediction on the ASCERTAIN dataset. Personal models performed differently depending on a dataset, but in most cases, we noticed lower error rates achieved when predicting non-standardized annotations paired with higher gain values (due to lower baseline RMSEs).

Depending on the dataset and task solved, we also saw differences in the preferred approach to the experimental setup used. In classification experiments, subject-wise models often achieved the highest accuracy and gain, followed by group-personalized models. On the other hand, in regression tasks, person-specific models never achieved the lowest error rates, and group-personalized models, although they performed best in terms of RMSE, were often paired with the lowest baselines (resulting in high gain values). When comparing gain values for statistically significant differences, we often found no differences between models that seemed to differ based solely on raw RMSE values or found differences between models that achieved similar error rates.

The above results provide preliminary evidence against our claims that personalized standardization could improve the results of models for emotion recognition from ECG signal. However, we would like to highlight the fact that results seem to depend on (1) a used dataset, (2) a solved task, and (3) an assumed experimental design. Also, in theory, proposed methods should perform better on information-rich real-life data, especially in rare situations when very intense emotions occur. Unfortunately, we could only test these methods on laboratory data due to the unavailability of a public real-life psychophysiology dataset and delays in processing our own dataset collected in everyday life.

*Chapter 10***CONCLUSIONS**

Throughout this dissertation, we introduced the reader to affective computing, presented the current state of the literature on emotion recognition methods for real life and on personalized affective computing, as well as described our own research in this field. In this last chapter, we summarize the key points of our work, provide conclusions regarding them, and reflect on the limitations of our research and possible future work.

10.1 Affect recognition literature

One of the major achievements described in this dissertation is the critical study of affective computing literature, with a special focus on the methods for emotion recognition. Not only did we describe methods used in emotion recognition for real life, but also similarities and differences between experiments conducted inside and outside the laboratory by comparing them across several identified study components, Chap. 3.

Our review revealed that it is necessary to conduct more studies on emotion psychophysiology in real life, but also that this topic attracts more and more scientists – with time, the number of articles focusing on everyday life increases, Sec. 3.1. However, as in almost every year within the review scope (except 2017), in-the-laboratory studies dominated over the in-the-field ones, affective studies still need to shift focus from controlled laboratory conditions to real life.

We also identified the main differences between laboratory and field studies across the characteristics of emotional experience and inductions, possible labeling and self-assessment methods, characteristics of utilized measuring devices and collected signals, and other constraints such as environment, required training, or amount of data collected. As a result, new challenges introduced by real-life studies became apparent, among others, (1) the necessity to collect information about context influencing emotions and utilize novel annotation methods, (2) designing machine learning models to consider multiple possible emotional states at the same time, (3) using methods that can learn from vast amounts of unlabeled data, (4) finding ways to combine subject-specific information with population-wise knowledge, (5)

designing novel validation methods, and (6) developing more accurate measuring devices. Moreover, if researchers do not contribute to open science, the whole affective computing field will suffer from a lack of data and code to develop new solutions and verify created methods.

We also contributed to literature research by reviewing articles that utilized personalization in affect recognition, Chap. 4. We identified four stages at which personalization can be introduced into developed methods, i.e., (1) choosing the subject of affect recognition, (2) designing the procedure used to create reasoning models, (3) designing algorithms and machine learning models, and (4) defining the tasks being solved. Additionally, we listed signals that are used in personalized affective computing and validation procedures assumed by researchers. Although personalization of affect recognition seems feasible, much work needs to be done in researching the extent to which developed solutions should be personalized, i.e., how much personal and general data to use. Also, we have yet to fully understand the impact that procedures have on results and proper ways of validating them, e.g., strategies or metrics.

10.2 Collecting data for emotion research

We also described the process of designing and executing a large everyday-life study and collecting a dataset containing everyday physiological and behavioral signal recordings annotated with emotional states, daily affect measures, and contextual data. As it contains a large amount of per-person data collected in everyday life, it may be useful for developing algorithms and methods for real-life reasoning, developing and analyzing personalization strategies, or studying daily-life patterns. However, collecting the LarField dataset was a difficult endeavor, comprising many months of preparation, testing, and designing, together with the engagement of the entire Emognition team.

As we had no prior experience conducting data collection experiments, starting with a much simpler experiment with controlled conditions, namely collecting the Emognition dataset, Chap. 5 was a good choice. It allowed us to test devices for recording physiological measures, discover issues with them (e.g., connection reliability, amount of noise and signal quality, the convenience of wearing them), and decide on the ones we wanted to use in further experiments. Due to the unsatisfactory performance of utilized devices, we included an ECG chest strap in the later experiments. Also, we learned a lot regarding the management and com-

munication with experiment subjects, including possible problems with people not following procedures, last-minute cancellations, and proper ways of communicating with humans unfamiliar with the topic.

Since emotions are sparse in everyday life, every captured emotional event is meaningful. Therefore, to increase the likelihood of measuring high-intensity emotional reactions, we utilized personalized machine learning models to recognize intense emotions in real-time and trigger self-assessment questionnaires. We researched different strategies for creating such models, especially personalized ones, and overcoming the cold start problem in affective studies, Chap 6. However, all machine learning models require large amounts of training samples, and collecting such amounts separately for each person may take a lot of time. Therefore, we proposed utilizing per-group personalization before the necessary quantity of per-person cases is reached.

10.3 Personalization for real-life affect recognition

Another problem that we delved into was the affect recognition in real life, with a particular focus on emotions. During literature research, we discovered that although researchers agree that people differ in their perception of affect, most scientific papers ignore this subjectivity and concentrate on general solutions. Simultaneously, the same studies often utilize poor experimental procedures, especially in model training and validation, thus potentially leading to knowledge leaks between training and testing data. This issue is well-presented by many scientists who develop their general models on the same set of participants that they are tested on. It results in a group-personalized setup instead of a universal one.

In our research, we focused on personalized methods for affect recognition. We confirmed that training models on personal data usually resulted in better recognition quality than creating subject-independent models while also outperforming baselines. Moreover, we confirmed that utilizing subject-specific or context-aware models results in the best recognition performance across different affective tasks. It is especially true for classification (Chaps. 7, 6, 9), but our investigation also demonstrated some potential in solving regression problems (Chaps. 7, 8, and 9). It is worth noting that regression of affective states is by itself a difficult problem, and the minority of experiments provided results better than the average baseline, Chap. 8. Nevertheless, our experimental studies proved that, to a certain extent, the affect can be recognized from physiological signals and that personalization

improves the recognition accuracy.

Unfortunately, we could not draw far-reaching conclusions regarding personalized data standardization for emotion recognition, Chap. 9. None of the designed procedures impacted results in a consistent way, neither when applied to input signals nor when used to process the annotations. Any differences that we were able to discern seem to be dataset-dependent, but we failed to identify significant factors governing them. However, the rationale behind the designed methods suggests that they might have some impact when tested on data collected in real life; then, we would expect more between-subject differences than in standardized environments. Regrettably, due to delays in collecting and processing our own dataset, unavailability of other datasets collected in daily life, and the approaching deadline for dissertation submission, we were unable to test the two-fold personalization methods on real-life data extensively. It will be the subject of our future work.

10.4 Limitations of this work and affective computing research

The presented research is not devoid of limitations, which could have affected the obtained results. One of them is the aforementioned delays in collecting and processing the LarField dataset. Although our results show that subject-dependent models, in most cases, outperform subject-independent ones, there is still room for improvement. Because data was not processed until June 2024, we were unable to perform more sophisticated experiments on the LarField dataset, deeply investigate relationships present in the data or thoroughly examine models' behavior.

Another limitation that affected the models' performance was the availability of resources for developing them. It is probable that the amount of available data, especially the number of per-person instances, was too low to train the personalized models properly, Chap. 9. We used a pre-trained representation learning model in our research, which should require fewer samples than training solely in a supervised manner. However, the number of training instances could still be too low to adjust it.

Problems with the availability of resources also affected the experiments on the LarField dataset, Chap. 7. Although the dataset contains many samples, most of them are unlabeled. Because of limited time and computational resources, we could only employ classical machine learning algorithms, which can be trained quickly and do not require vast numbers of instances to capture patterns in data. However, those models were not designed to contain vast amounts of knowledge, thus leading

to them possibly capturing only a tiny subset of possible relationships between features.

Moreover, as we utilized simple machine learning models, we had to perform very restrictive data cleaning. It resulted in the discarding of many data instances. Therefore, we could not split data into separate training and testing sets and had to carry out cross-validation testing on all data. We decided to avoid overfitting at all costs and, therefore, had to resign from doing parameter optimization for utilized algorithms. Without using a separate test set for the final evaluation, any such procedure introduces knowledge leaks¹, resulting in overfitted models and overestimated values of performance metrics. Therefore, although the models were not as accurate as they could, the reached accuracy estimates should correctly reflect their capabilities in recognizing everyday affect.

10.5 Summary and future work

Even though personalized affective computing has yet to become widely adapted, other authors proposed some personalization methods prior to this dissertation, Chap. 4. In our own studies, we pushed the research further by:

1. confirming the benefits of utilizing personalized models; both for groups of people and for individuals, Chaps. 6 - 9;
2. investigating the amount of data required to improve predictions over non-personalized training, Chap. 6;
3. studying the impact of context (Chap. 7) and processing (Chap. 9) while creating the personalized models;
4. examining different affective tasks that can be modeled using machine learning methods (Chaps. 6 - 9) and analyzing of properties of such models.

Apart from the above contributions to personalized methods for affective computing, our work provides some other benefits to the field, such as

1. a critical literature review of methods for real-life emotion recognition, Chap. 3;
2. a review of methods used in personalized affective computing, Chap. 4;

¹They are introduced by choosing specific values of different parameters to best fit the data, therefore biasing the solution.

3. a publicly-available dataset of physiological measurements collected with unobtrusive off-the-shelf wearables, and their emotional annotations, Chap. 5;
4. a large information-rich dataset collected in real life (Chap. 7) that we plan to publish soon;
5. a description and results of using a *Big Team Science* competition as a method of researching difficult research questions, Chap. 8;
6. conclusions regarding commensurability, replicability, and comparability in affective computing studies utilizing machine learning methods, Chaps. 3, 4 and 8.

Continuation of research presented in this dissertation may take several forms. As the research field is constantly growing, periodic updates to literature studies are needed to be up to date with the newest developments in wearable devices, methods for recognizing affect, available datasets, and procedures for conducting studies and experiments. The developed methods of personalization via data standardization, although they did not significantly impact the recognition quality on laboratory data, could reveal their potential on a dataset collected outside the laboratory. In such a wild environment, we would expect more between-subject differences and more within-subject similarities. Therefore, we are planning to run the experiments on the LarField dataset right after the ECG signals are satisfactorily synchronized with the rest of the data.

Regarding the experiments on the collected LarField dataset, its information-rich nature allows for studying various states that a person may find themselves in during their everyday life. In addition to daily stress, mood, health, and emotion indicators, which we studied briefly in this dissertation, one can explore many other aspects of people's daily lives and their impact on different measures, such as (1) stress levels, its changes over time, and how they depend on other collected measures, (2) correlations between different self-reported affect measures, (3) cycles in people's self-reported affect and other collected measures, or (4) impact of physical activity and daily mobility on self-reported affect. Not only do we plan to carry out more research on this dataset, but we also will publish the dataset once the data is thoroughly investigated, cleansed, synchronized, and described.

One may also investigate the extent to which the links between emotions, physiology, and behavior are subjective and search for some general patterns in their activation.

The fact that not only subject-specific training but also group-personalized models often outperformed baselines in our experiments may suggest the presence of both types of relationships. The overall good performance of group-personalized models could stem from data sparsity in high-dimensional feature space. We can test whether models learn the general emotion-physiology links or just simple averages for each person. Researching this problem while taking into account all possible aspects impacting emotion recognition would require designing and conducting an entirely new study, as none of the datasets and methods presented in the dissertation are appropriate for addressing such a question directly.

Overall, this dissertation is but a small contribution to emotion and affective research compared to all the work that still needs to be done. Nevertheless, our experiments suggest that scientists should put much more effort into investigating the subjectivity of affect, factors impacting it, and methods for personalizing machine learning models.

BIBLIOGRAPHY

- [1] D. Antonio, “Descartes’ error,” *Emotion, reason and the human brain*, 1994.
- [2] H. A. Simon, “Motivational and emotional controls of cognition.,” *Psychological review*, vol. 74, no. 1, p. 29, 1967.
- [3] D. Kahneman, *Thinking, fast and slow*. Macmillan, 2011.
- [4] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, “Emotion and decision making,” *Annual review of psychology*, vol. 66, pp. 799–823, 2015.
- [5] T. Telford, ““emotion detection’ ai is a \$20 billion industry. research says it can’t do what it claims,” *The Washington Post*, 2019.
- [6] I. M. I. Research, *Emotion Detection And Recognition Market Size, Growth Forecast- Industry Outlook to 2028*, Jun. 2022. [Online]. Available: <https://www.intellectualmarketinsights.com/report/emotion-detection-and-recognition-market-size/imi-000140>.
- [7] P. Ekman, R. W. Levenson, and W. V. Friesen, “Autonomic nervous system activity distinguishes among emotions,” *science*, vol. 221, no. 4616, pp. 1208–1210, 1983.
- [8] S. D. Kreibig, “Autonomic nervous system activity in emotion: A review,” *Biological psychology*, vol. 84, no. 3, pp. 394–421, 2010.
- [9] M. Behnke, S. D. Kreibig, L. D. Kaczmarek, M. Assink, and J. J. Gross, “Autonomic nervous system activity during positive emotions: A meta-analytic review,” *Emotion Review*, vol. 14, no. 2, pp. 132–160, 2022.
- [10] R. W. Picard, *Affective computing*. MIT press, 2000.
- [11] 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*, 2019, pp. 211–219.
- [12] P. Schmidt, A. Reiss, R. Dürichen, and K. Van Laerhoven, “Wearable-based affect recognition—a review,” *Sensors*, vol. 19, no. 19, p. 4079, 2019.
- [13] L. Barrett, “Debate about universal facial expressions goes big,” *Nature*, vol. 589, no. 7841, pp. 202–203, 2021.
- [14] A. Wierzbicka, “Human emotions: Universal or culture-specific?” *American anthropologist*, vol. 88, no. 3, pp. 584–594, 1986.
- [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 *Int. Conf. Affect. Comput. Intell. Interact. (ACII 2021)*, IEEE, 2021, pp. 1–3.

- [16] S. Saganowski, B. Perz, A. G. Polak, and P. Kazienko, “Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review,” *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1876–1897, 2023. DOI: 10.1109/TAFFC.2022.3176135.
- [17] D. C. Ong, “An ethical framework for guiding the development of affectively-aware artificial intelligence,” in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2021, pp. 1–8.
- [18] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Polak, “Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements,” *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.
- [19] D. Cutuli, “Cognitive reappraisal and expressive suppression strategies role in the emotion regulation: An overview on their modulatory effects and neural correlates,” *Frontiers in systems neuroscience*, vol. 8, p. 110 157, 2014.
- [20] L. F. Barrett, *How Emotions Are Made: The Secret Life of the Brain*. Mariner Books, 2017.
- [21] L. F. Barrett, “Variety is the spice of life: A psychological construction approach to understanding variability in emotion,” *Cognition and emotion*, vol. 23, no. 7, pp. 1284–1306, 2009.
- [22] A. Wierzbicka, *Emotions across Languages and Cultures: Diversity and Universals*. Cambridge University Press, 1999.
- [23] N. Ramzan and N. Amjad, “Cross cultural variation in emotion regulation: A systematic review,” *Annals of King Edward Medical University*, vol. 23, no. 1, 2017.
- [24] S. Saganowski, J. Komoszyńska, M. Behnke, B. Perz, D. Kunc, B. Klich, Ł. D. Kaczmarek, and P. Kazienko, “Emognition dataset: Emotion recognition with self-reports, facial expressions, and physiology using wearables,” *Scientific data*, vol. 9, no. 1, p. 158, 2022. DOI: 10.1038/s41597-022-01262-0.
- [25] J. Komoszyńska, D. Kunc, B. Perz, A. Hebko, P. Kazienko, and S. Saganowski, “Designing and executing a large-scale real-life affective study,” in *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, IEEE, 2024, pp. 505–510.
- [26] D. Kunc, J. Komoszyńska, B. Perz, S. Saganowski, and P. Kazienko, “Emognition system-wearables, physiology, and machine learning for real-life emotion capturing,” in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, IEEE, 2023, pp. 1–3.

- [27] D. Kunc, J. Komoszyńska, B. Perz, P. Kazienko, and S. Saganowski, “Real-life validation of emotion detection system with wearables,” in *International Work-Conference on the Interplay Between Natural and Artificial Computation*, Springer, 2022, pp. 45–54.
- [28] S. Saganowski, D. Kunc, B. Perz, J. Komoszyńska, M. Behnke, and P. Kazienko, “The cold start problem and per-group personalization in real-life emotion recognition with wearables,” in *2022 IEEE Int. Conf. Pervasive Comput. Commun. Workshops, WristSense 2022 - The Eighth Workshop on Sensing Systems and Applications Using Wrist Worn Smart Devices*, Best Paper Award, IEEE, 2022, pp. 812–817. doi: 10.1109/PerComWorkshops53856.2022.9767233.
- [29] K. Avramidis, D. Kunc, B. Perz, K. Adsul, T. Feng, P. Kazienko, S. Saganowski, and S. Narayanan, “Scaling representation learning from ubiquitous ecg with state-space models,” *IEEE Journal of Biomedical and Health Informatics*, 2024.
- [30] B. Perz, *Introduction to the epic challenge*, 11th International Conference on Affective Computing & Intelligent Interaction (ACII), MIT Media Lab, Cambridge, MA, USA, unpublished presentation, Sep. 2023.
- [31] B. Perz, *The emotion physiology and experience collaboration (epic) challenge*, 2023 Big Team Science Conference (BTSCON), online, unpublished presentation, Oct. 2023.
- [32] N. A. Coles, B. Perz, M. Behnke, J. C. Eichstaedt, S.-H. Kim, T. N. Vu, C. Raman, J. Tejada, G. Zhang, T. Cui, S. Podder, R. Chavda, S. Pandey, A. Upadhyay, J. I. Padilla-Buritica, C. J. Barrera Causil, L. Ji, F. Dollack, K. Kiyokawa, H. Liu, M. Perusquia-Hernandez, H. Uchiyama, X. Wei, H. Cao, Z. Yang, A. Iancarelli, K. McVeigh, Y. Wang, I. M. Berwian, J. C. Chiu, M. Dan-Mircea, E. C. Nook, H. I. Vartiainen, C. Whiting, Y. Won Cho, S.-M. Chow, Z. F. Fisher, Y. Li, X. Xiong, Y. Shen, E. Tagliazucchi, L. Bugnon, R. Ospina, N. M. Bruno, T. A. D’Amelio, F. Zamberlan, L. R. Mercado Diaz, J. O. Pinzon-Arenas, H. F. Posada-Quintero, M. Bilalpur, S. Hinduja, F. Marmolejo-Ramos, S. Canavan, L. Jivnani, and S. Saganowski, “Big team science reveals promises and limitations of machine learning efforts to model the physiological basis of affective experience,” *Nature Human Behaviour*, 2024, In reviews.
- [33] K. Johnson and L. Tze Lim, *Caltech Thesis LaTeX Template (without logo)*, 2016. [Online]. Available: <https://www.overleaf.com/latex/templates/caltech-thesis-latex-template-without-logo/xsjkwzcftrym>.
- [34] P. Desmet, *Designing emotions*, 2002.
- [35] R. Jhangiani and H. Tarry, “Principles of social psychology-1st international h5p edition,” 2022.

- [36] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant.," *Journal of personality and social psychology*, vol. 76, no. 5, p. 805, 1999.
- [37] L. F. Barrett and E. Bliss-Moreau, "Affect as a psychological primitive," *Advances in experimental social psychology*, vol. 41, pp. 167–218, 2009.
- [38] N. Schneiderman, G. Ironson, and S. D. Siegel, "Stress and health: Psychological, behavioral, and biological determinants," *Annu. Rev. Clin. Psychol.*, vol. 1, pp. 607–628, 2005.
- [39] P. Ekman and W. V. Friesen, *Facial action coding system: Investigator's guide*. Consulting Psychologists Press, 1978.
- [40] C. E. Izard, *Human Emotions*. New York: Springer Science & Business Media, 1977, ISBN: 978-1-4899-2211-3.
- [41] R. Lazarus, *Psychological stress and the coping process*, 1966.
- [42] M. E. 1735, *Plutchik-wheel.svg*, 2011. [Online]. Available: <https://en.wikipedia.org/wiki/File:Plutchik-wheel.svg>.
- [43] mrAnmol, *Circumplex model of emotion.svg*, 2023. [Online]. Available: https://en.m.wikipedia.org/wiki/File:Circumplex_model_of_emotion.svg.
- [44] S. Buechel and U. Hahn, "Emotion analysis as a regression problem—dimensional models and their implications on emotion representation and metrical evaluation," in *ECAI 2016*, IOS Press, 2016, pp. 1114–1122.
- [45] R. Plutchik, *Emotions and life: Perspectives from psychology, biology, and evolution*. American Psychological Association, 2003.
- [46] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.
- [47] A. Mehrabian, "Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies," 1980.
- [48] P. Ekman, "Facial expression and emotion.," *American psychologist*, vol. 48, no. 4, p. 384, 1993.
- [49] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, 2020.
- [50] S. Zhao, G. Jia, J. Yang, G. Ding, and K. Keutzer, "Emotion recognition from multiple modalities: Fundamentals and methodologies," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 59–73, 2021.
- [51] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *sensors*, vol. 18, no. 2, p. 401, 2018.

- [52] M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” *Speech Communication*, vol. 116, pp. 56–76, 2020.
- [53] J. Z. Lim, J. Mountstephens, and J. Teo, “Emotion recognition using eye-tracking: Taxonomy, review and current challenges,” *Sensors*, vol. 20, no. 8, p. 2384, 2020.
- [54] P. Nandwani and R. Verma, “A review on sentiment analysis and emotion detection from text,” *Social network analysis and mining*, vol. 11, no. 1, p. 81, 2021.
- [55] R. Fernandez and R. W. Picard, “Signal processing for recognition of human frustration,” in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP’98 (Cat. No. 98CH36181)*, IEEE, vol. 6, 1998, pp. 3773–3776.
- [56] M. B. H. Wiem and Z. Lachiri, “Emotion classification in arousal valence model using mahnob-hci database,” *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 3, 2017.
- [57] W. Chang, L. Xu, Q. Yang, and Y. Ma, “Eeg signal-driven human–computer interaction emotion recognition model using an attentional neural network algorithm,” *Journal of Mechanics in Medicine and Biology*, vol. 23, no. 08, p. 2340080, 2023.
- [58] G. J. Nalepa, K. Kutt, B. Giżycka, P. Jemiolo, and S. Bobek, “Analysis and use of the emotional context with wearable devices for games and intelligent assistants,” *Sensors*, vol. 19, no. 11, p. 2509, 2019.
- [59] T. Xu, R. Yin, L. Shu, and X. Xu, “Emotion recognition using frontal eeg in vr affective scenes,” in *2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC)*, IEEE, vol. 1, 2019, pp. 1–4.
- [60] K. Kutt, G. J. Nalepa, B. Giżycka, P. Jemiolo, and M. Adamczyk, “Bandreader—a mobile application for data acquisition from wearable devices in affective computing experiments,” in *2018 11th International Conference on Human System Interaction (HSI)*, IEEE, 2018, pp. 42–48.
- [61] J. Rubin, H. Eldardiry, R. Abreu, S. Ahern, H. Du, A. Pattekar, and D. G. Bobrow, “Towards a mobile and wearable system for predicting panic attacks,” in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp ’15, Osaka, Japan: Association for Computing Machinery, 2015, pp. 529–533, ISBN: 9781450335744. DOI: 10.1145/2750858.2805834. [Online]. Available: <https://doi.org/10.1145/2750858.2805834>.
- [62] A. Saeed, T. Ozcelebi, J. Lukkien, J. B. van Erp, and S. Trajanovski, “Model adaptation and personalization for physiological stress detection,” in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018, pp. 209–216. DOI: 10.1109/DSAA.2018.00031.

- [63] M.-S. Dao, D.-T. Dang-Nguyen, A. Kasem, and H. Tran-The, “Healthy-classroom - 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 Recognit App. and Methods (ICPRAM)*, Scitepress–Science and Technology Publications, 2018, pp. 685–691.
- [64] D. Wu, X. Han, Z. Yang, and R. Wang, “Exploiting transfer learning for emotion recognition under cloud-edge-client collaborations,” *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 2, pp. 479–490, 2021. doi: 10.1109/JSAC.2020.3020677.
- [65] H. Yu and A. Sano, “Passive sensor data based future mood, health, and stress prediction: User adaptation using deep learning,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 5884–5887. doi: 10.1109/EMBC44109.2020.9176242.
- [66] K. D. Bartl-Pokorny, M. Pykała, P. Uluer, D. E. Barkana, A. Baird, H. Kose, T. Zorcec, B. Robins, B. W. Schuller, and A. Landowska, “Robot-based intervention for children with autism spectrum disorder: A systematic literature review,” *IEEE Access*, vol. 9, pp. 165 433–165 450, 2021.
- [67] A. Landowska, A. Karpus, T. Zawadzka, B. Robins, D. Erol Barkana, H. Kose, T. Zorcec, and N. Cummins, “Automatic emotion recognition in children with autism: A systematic literature review,” *Sensors*, vol. 22, no. 4, p. 1649, 2022.
- [68] D. Antos and A. Pfeffer, “Using emotions to enhance decision-making,” in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.
- [69] M. Behnke, S. Saganowski, D. Kunc, and P. Kazienko, “Ethical considerations and checklist for affective research with wearables,” *IEEE Transactions on Affective Computing*, 2022.
- [70] J. Hernandez, J. Lovejoy, D. McDuff, J. Suh, T. O’Brien, A. Sethumadhavan, G. Greene, R. Picard, and M. Czerwinski, “Guidelines for assessing and minimizing risks of emotion recognition applications,” in *2021 9th International conference on affective computing and intelligent interaction (ACII)*, IEEE, 2021, pp. 1–8.
- [71] F. Cabitza, A. Campagner, and M. Mattioli, “The unbearable (technical) unreliability of automated facial emotion recognition,” *Big data & society*, vol. 9, no. 2, p. 20 539 517 221 129 549, 2022.
- [72] H. Schlosberg, “Three dimensions of emotion.,” *Psychological review*, vol. 61, no. 2, p. 81, 1954.

- [73] R. W. Levenson, P. Ekman, and W. V. Friesen, "Voluntary facial action generates emotion-specific autonomic nervous system activity," *Psychophysiology*, vol. 27, no. 4, pp. 363–384, 1990.
- [74] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004.
- [75] J. Hayano, T. Tanabiki, S. Iwata, K. Abe, and E. Yuda, "Estimation of emotions by wearable biometric sensors under daily activities," in *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, IEEE, 2018, pp. 240–241.
- [76] 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*, ser. UbiComp '16, Heidelberg, Germany: Association for Computing Machinery, 2016, pp. 1153–1161, ISBN: 9781450344623. DOI: 10.1145/2968219.2968302. [Online]. Available: <https://doi.org/10.1145/2968219.2968302>.
- [77] A. Fortin-Côté, N. Beaudin-Gagnon, A. Campeau-Lecours, S. Tremblay, and P. L. Jackson, "Affective computing out-of-the-lab: The cost of low cost," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, IEEE, 2019, pp. 4137–4142.
- [78] D. Pollreisz and N. TaheriNejad, "A simple algorithm for emotion recognition, using physiological signals of a smart watch," in *Proc IEEE Eng Med Biol Soc (EMBC)*, IEEE, 2017, pp. 2353–2356.
- [79] B. Zhao, Z. Wang, Z. Yu, and B. Guo, "Emotionsense: Emotion recognition based on wearable wristband," in *SmartWorld, Ubiqu. Intell., Internet of People and Smart City Innovation*, IEEE, 2018, pp. 346–355.
- [80] A. Albraikan, D. P. Tobón, and A. El Saddik, "Toward user-independent emotion recognition using physiological signals," *IEEE Sensors Journal*, vol. 19, no. 19, pp. 8402–8412, 2018.
- [81] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran, "Long short term memory hyperparameter optimization for a neural network based emotion recognition framework," *IEEE Access*, vol. 6, pp. 49 325–49 338, 2018.
- [82] M. Ragot, N. Martin, S. Em, N. Pallamin, and J.-M. Diverrez, "Emotion recognition using physiological signals: Laboratory vs. wearable sensors," in *Advances in Human Factors in Wearable Technologies and Game Design*, Springer, 2018, pp. 15–22.
- [83] L. Romeo, A. Cavallo, L. Pepa, N. Berthouze, and M. Pontil, "Multiple instance learning for emotion recognition using physiological signals," *IEEE Transactions on Affective Computing*, 2019.

- [84] R. Wampfler, S. Klingler, B. Solenthaler, V. Schinazi, and M. Gross, “Affective state prediction in a mobile setting using wearable biometric sensors and stylus,” in *Proceedings of The 12th International Conf. on Educational Data Mining (EDM 2019)*, 2019, pp. 198–207.
- [85] T. Martens, M. Niemann, and U. Dick, “Sensor measures of affective leaning,” *Front. Psychol.*, vol. 11, 2020.
- [86] P. Saxena, S. Dabas, D. Saxena, N. Ramachandran, and S. I. Ahamed, “Reconstructing compound affective states using physiological sensor data,” in *2020 IEEE 44th Annu. Comp., Softw., and Appl. Conf.*, IEEE, 2020, pp. 1241–1249.
- [87] F. Nasoz, O. Ozyer, C. L. Lisetti, and N. Finkelstein, “Multimodal affective driver interfaces for future cars,” in *Proc ACM Int Conf Multimed*, ACM, 2002, pp. 319–322.
- [88] L. Hu, J. Yang, M. Chen, Y. Qian, and J. J. Rodrigues, “Scal-svsc: Smart clothing for effective interaction with a sustainable vital sign collection,” *Fut Gen Comp Sys*, vol. 86, pp. 329–338, 2018.
- [89] C. L. Lisetti and F. Nasoz, “Using noninvasive wearable computers to recognize human emotions from physiological signals,” *EURASIP J. Adv. Signal Process.*, vol. 2004, pp. 1672–1687, 2004.
- [90] C. L. Lisetti and F. Nasoz, “Categorizing autonomic nervous system (ans) emotional signals using bio-sensors for hri within the maui paradigm,” in *ROMAN 2006 - IEEE Int. Symp. Robot Hum. Interact. Commun.*, IEEE, 2006, pp. 277–284.
- [91] A. F. Bulagang, J. Mountstephens, and J. T. T. Wi, “Tuning support vector machines for improving four-class emotion classification in virtual reality (vr) using heart rate features,” in *J. Phys. Conf. Ser.*, IOP Publishing, vol. 1529, 2020, p. 052 069.
- [92] L. Shu, Y. Yu, W. Chen, H. Hua, Q. Li, J. Jin, and X. Xu, “Wearable emotion recognition using heart rate data from a smart bracelet,” *Sensors*, vol. 20, no. 3, p. 718, 2020.
- [93] F. Setiawan, A. G. Prabono, S. A. Khowaja, W. Kim, K. Park, B. N. Yahya, S.-L. Lee, and J. P. Hong, “Fine-grained emotion recognition: Fusion of physiological signals and facial expressions on spontaneous emotion corpus,” *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 35, no. 3, pp. 162–178, 2020.
- [94] A. J. Majumder, J. W. Dedmond, S. Jones, and A. A. Asif, “A smart cyber-human system to support mental well-being through social engagement,” in *2020 IEEE 44th Annu. Comp., Softw., and Appl. Conf.*, IEEE, 2020, pp. 1050–1058.

- [95] K. Rattanyu, M. Ohkura, and M. Mizukawa, "Emotion monitoring from physiological signals for service robots in the living space," in *ICCAS 2010*, IEEE, 2010, pp. 580–583.
- [96] K. Rattanyu and M. Mizukawa, "Emotion recognition using biological signal in intelligent space," in *International Conference on Human-Computer Interaction*, Springer, 2011, pp. 586–592.
- [97] H. W. Guo, Y. S. Huang, J. C. Chien, and J. S. Shieh, "Short-term analysis of heart rate variability for emotion recognition via a wearable ecg device," in *2015 Int. Conf. Intell. Inform. Biomed. Sci. (ICIIBMS)*, IEEE, 2015, pp. 262–265.
- [98] H.-W. Guo, Y.-S. Huang, C.-H. Lin, J.-C. Chien, K. Haraikawa, and J.-S. Shieh, "Heart rate variability signal features for emotion recognition by using principal component analysis and support vectors machine," in *2016 IEEE 16th Int. Conf. Bioinform. Bioeng. (BIBE)*, IEEE, 2016, pp. 274–277.
- [99] 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, e10153, 2018.
- [100] 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.
- [101] A. Albraikan, B. Hafidh, and A. El Saddik, "Iaware: A real-time emotional biofeedback system based on physiological signals," *IEEE Access*, vol. 6, pp. 78 780–78 789, 2018.
- [102] 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.
- [103] 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.
- [104] E. Kanjo, E. M. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Information Fusion*, vol. 49, pp. 46–56, 2019.
- [105] S. Kim, K. Patra, A. Kim, K.-P. Lee, A. Segev, and U. Lee, "Sensors know which photos are memorable," in *Proc. 2017 CHI Conf. Ext. Abstr. Hum. Factors in Comput. Syst.*, 2017, pp. 2706–2713.
- [106] N. T. Nguyen, N. V. Nguyen, M. H. T. Tran, and B. T. Nguyen, "A potential approach for emotion prediction using heart rate signals," in *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, IEEE, 2017, pp. 221–226.

- [107] Y. Kadoya, M. S. R. Khan, S. Watanapongvanich, and P. Binnagan, “Emotional status and productivity: Evidence from the special economic zone in laos,” *Sustainability*, vol. 12, no. 4, p. 1544, 2020.
- [108] F. Setiawan, S. A. Khowaja, A. G. Prabono, B. N. Yahya, and S.-L. Lee, “A framework for real time emotion recognition based on human ans using pervasive device,” in *2018 IEEE 42nd Annu. Comp., Softw., and Appl. Conf.*, IEEE, vol. 1, 2018, pp. 805–806.
- [109] M. Timmers, A. Fischer, and A. Manstead, “Ability versus vulnerability: Beliefs about men’s and women’s emotional behaviour,” *Cognition and emotion*, vol. 17, no. 1, pp. 41–63, 2003.
- [110] M. D. Robinson and J. T. Johnson, “Is it emotion or is it stress? gender stereotypes and the perception of subjective experience,” *Sex Roles*, vol. 36, no. 3-4, pp. 235–258, 1997.
- [111] L. R. Brody and J. A. Hall, “Gender and emotion in context,” *Handbook of emotions*, vol. 3, pp. 395–408, 2008.
- [112] L. Fernández-Aguilar, J. Ricarte, L. Ros, and J. M. Latorre, “Emotional differences in young and older adults: Films as mood induction procedure,” *Front. Psychol.*, vol. 9, p. 1110, 2018.
- [113] A. Costa, J. A. Rincon, C. Carrascosa, V. Julian, and P. Novais, “Emotions detection on an ambient intelligent system using wearable devices,” *Future Gen. Computer Systems*, vol. 92, pp. 479–489, 2019.
- [114] E. Y. Chan and S. J. Maglio, “Coffee cues elevate arousal and reduce level of construal,” *Consciousness and cognition*, vol. 70, pp. 57–69, 2019.
- [115] P. J. Green, R. Kirby, and J. Suls, “The effects of caffeine on blood pressure and heart rate: A review,” *Annals of Behavioral Medicine*, vol. 18, no. 3, pp. 201–216, 1996.
- [116] 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*, 2020, pp. 482–487.
- [117] A. A. Schuurmans, P. de Looff, K. S. Nijhof, C. Rosada, R. H. Scholte, A. Popma, and R. Otten, “Validity of the empatica e4 wristband to measure heart rate variability (hrv) parameters: A comparison to electrocardiography (ecg),” *Journal of medical systems*, vol. 44, no. 11, pp. 1–11, 2020.
- [118] A. Borrego, J. Latorre, M. Alcañiz, and R. Llorens, “Reliability of the empatica e4 wristband to measure electrodermal activity to emotional stimuli,” in *2019 International Conference on Virtual Rehabilitation (ICVR)*, IEEE, 2019, pp. 1–2.

- [119] P. Konstantinou, A. Trigeorgi, C. Georgiou, A. T. Gloster, G. Panayiotou, and M. Karekla, “Comparing apples and oranges or different types of citrus fruits? using wearable versus stationary devices to analyze psychophysiological data,” *Psychophysiology*, vol. 57, no. 5, e13551, 2020.
- [120] D. Krech, R. S. Crutchfield, and N. Livson, *Elements of psychology*, 3rd edition. New York: Alfred A. Knopf, 1974.
- [121] 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, p. 6535, 2020.
- [122] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller, “End-to-end learning for dimensional emotion recognition from physiological signals,” in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2017, pp. 985–990.
- [123] M. Feurer, A. Klein, K. Eggenberger, J. T. Springenberg, M. Blum, and F. Hutter, “Auto-sklearn: Efficient and robust automated machine learning,” in *Automated Machine Learning*, Springer, Cham, 2019, pp. 113–134.
- [124] L. F. Barrett, B. Mesquita, and M. Gendron, “Context in emotion perception,” *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286–290, 2011.
- [125] J. Choi, B. Ahmed, and R. Gutierrez-Osuna, “Development and evaluation of an ambulatory stress monitor based on wearable sensors,” *IEEE transactions on information technology in biomedicine*, vol. 16, no. 2, pp. 279–286, 2011.
- [126] S. Du, Y. Tao, and A. M. Martinez, “Compound facial expressions of emotion,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, E1454–E1462, 2014.
- [127] A. S. Cowen and D. Keltner, “Self-report captures 27 distinct categories of emotion bridged by continuous gradients,” *Proc. Acad. Nat. Sci.*, vol. 114, no. 38, E7900–E7909, 2017.
- [128] 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 Int. Conf. Pervasive Comput. Commun. Workshops*, IEEE, 2021, pp. 348–351.
- [129] W. Sasaki, J. Nakazawa, and T. Okoshi, “Comparing esm timings for emotional estimation model with fine temporal granularity,” in *Symp. Perv. Ubiq. Comp. Wear. Comp.*, ACM, 2018, pp. 722–725.
- [130] P. Denman, E. Lewis, S. Prasad, J. Healey, H. Syed, and L. Nachman, “Affsens: A mobile platform for capturing affect in context,” in *Proc. of the 20th Int. Conf. on Human-Computer Interaction with Mobile Devices and Services Adjunct*, ACM, 2018, pp. 321–326.

- [131] J. A. Rincon, A. Costa, P. Novais, V. Julian, and C. Carrascosa, “Intelligent wristbands for the automatic detection of emotional states for the elderly,” in *Int. Conf. Intell. Data Eng. Autom. Lear.*, Springer, 2018, pp. 520–530.
- [132] 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*, vol. 3, no. 1, pp. 18–31, 2011.
- [133] R. Plutchik, “A general psychoevolutionary theory of emotion,” in *Theories of emotion*, Elsevier, 1980, pp. 3–33.
- [134] L. F. Barrett, *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017.
- [135] N. Jaques, S. Taylor, A. Sano, and R. Picard, “Multi-task, multi-kernel learning for estimating individual wellbeing,” in *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*, vol. 898, 2015, p. 3.
- [136] F. Larradet, R. Niewiadomski, G. Barresi, D. G. Caldwell, and L. S. Mattos, “Toward emotion recognition from physiological signals in the wild: Approaching the methodological issues in real-life data collection,” *Frontiers in psychology*, vol. 11, p. 1111, 2020.
- [137] G. Stemmler and J. Wacker, “Personality, emotion, and individual differences in physiological responses,” *Biological psychology*, vol. 84, no. 3, pp. 541–551, 2010.
- [138] H. Yu, T. Vaessen, I. Myin-Germeys, and A. Sano, “Modality fusion network and personalized attention in momentary stress detection in the wild,” in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2021, pp. 1–8.
- [139] A. Sano, S. Taylor, A. W. McHill, A. J. Phillips, L. K. Barger, E. Klerman, and R. Picard, “Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: Observational study,” *Journal of medical Internet research*, vol. 20, no. 6, e210, 2018.
- [140] K. H. Cheah, H. Nisar, V. V. Yap, and C.-Y. Lee, “Short-time-span eeg-based personalized emotion recognition with deep convolutional neural network,” in *2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 2019, pp. 78–83. DOI: 10.1109/ICSIPA45851.2019.8977786.
- [141] K. Gupta, J. Lazarevic, Y. S. Pai, and M. Billinghurst, “Affectivelyvr: Towards vr personalized emotion recognition,” in *Proceedings of the 26th ACM Symposium on Virtual Reality Software and Technology*, 2020, pp. 1–3.

- [142] J. Tervonen, S. Puttonen, M. J. Sillanpää, L. Hopsu, Z. Homorodi, J. Keränen, J. Pajukanta, A. Tolonen, A. Lämsä, and J. Mäntyjärvi, “Personalized mental stress detection with self-organizing map: From laboratory to the field,” *Computers in Biology and Medicine*, vol. 124, p. 103 935, 2020, ISSN: 0010-4825. DOI: <https://doi.org/10.1016/j.combiomed.2020.103935>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0010482520302717>.
- [143] A. Muaremi, B. Arnrich, and G. Tröster, “Towards measuring stress with smartphones and wearable devices during workday and sleep,” *BioNanoScience*, vol. 3, no. 2, pp. 172–183, 2013.
- [144] G. Udovičić, J. Đerek, M. Russo, and M. Sikora, “Wearable emotion recognition system based on gsr and ppg signals,” in *Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care*, ser. MMHealth '17, Mountain View, California, USA: Association for Computing Machinery, 2017, pp. 53–59, ISBN: 9781450355049. DOI: 10.1145/3132635.3132641. [Online]. Available: <https://doi.org/10.1145/3132635.3132641>.
- [145] G. Valenza, L. Citi, A. Lanatá, E. P. Scilingo, and R. Barbieri, “Revealing real-time emotional responses: A personalized assessment based on heart-beat dynamics,” *Scientific reports*, vol. 4, no. 1, pp. 1–13, 2014.
- [146] M. Zhao, F. Adib, and D. Katabi, “Emotion recognition using wireless signals,” in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, 2016, pp. 95–108.
- [147] R. K. Sah and H. Ghasemzadeh, “Stress classification and personalization: Getting the most out of the least,” *arXiv preprint arXiv:2107.05666*, 2021.
- [148] H. LOTFALINEZHAD and A. Maleki, “Application of multiscale fuzzy entropy features for multilevel subject-dependent emotion recognition,” *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 27, no. 6, pp. 4070–4081, 2019.
- [149] F. I. Indikawati and S. Winiarti, “Stress detection from multimodal wearable sensor data,” in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, vol. 771, 2020, p. 012 028.
- [150] J. Hernandez, R. R. Morris, and R. W. Picard, “Call center stress recognition with person-specific models,” in *International Conference on Affective Computing and Intelligent Interaction*, Springer, 2011, pp. 125–134.
- [151] A. Dessai and H. Virani, “Emotion classification based on cwt of ecg and gsr signals using various cnn models,” *Electronics*, vol. 12, no. 13, p. 2795, 2023.

- [152] E. Mattern, R. R. Jackson, R. Doshmanziari, M. Dewitte, D. Varagnolo, and S. Knorn, "Emotion recognition from physiological signals collected with a wrist device and emotional recall," *Bioengineering*, vol. 10, no. 11, p. 1308, 2023.
- [153] S. D. ArulDass and P. Jayagopal, "Identifying complex emotions in alexithymia affected adolescents using machine learning techniques," *Diagnostics*, vol. 12, no. 12, p. 3188, 2022.
- [154] H. Yu, E. B. Klerman, R. W. Picard, and A. Sano, "Personalized wellbeing prediction using behavioral, physiological and weather data," in *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, 2019, pp. 1–4. DOI: 10.1109/BHI.2019.8834456.
- [155] B. Li and A. Sano, "Early versus late modality fusion of deep wearable sensor features for personalized prediction of tomorrow's mood, health, and stress*," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 5896–5899. DOI: 10.1109/EMBC44109.2020.9175463.
- [156] B. Li and A. Sano, "Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 2, Jun. 2020. DOI: 10.1145/3397318. [Online]. Available: <https://doi.org/10.1145/3397318>.
- [157] I. O. Joudeh, A.-M. Cretu, S. Bouchard, and S. Guimond, "Prediction of continuous emotional measures through physiological and visual data," *Sensors*, vol. 23, no. 12, p. 5613, 2023.
- [158] K. Yang, B. Tag, Y. Gu, C. Wang, T. Dingler, G. Wadley, and J. Goncalves, "Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022, pp. 562–570.
- [159] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, "Corrnet: Fine-grained emotion recognition for video watching using wearable physiological sensors," *Sensors*, vol. 21, no. 1, p. 52, 2020.
- [160] S. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 200–213, 2020. DOI: 10.1109/TAFFC.2017.2784832.
- [161] Z. Tian, D. Huang, S. Zhou, Z. Zhao, and D. Jiang, "Personality first in emotion: A deep neural network based on electroencephalogram channel attention for cross-subject emotion recognition," *Royal Society Open Science*, vol. 8, no. 8, p. 201976, 2021. DOI: 10.1098/rsos.201976.

- [162] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva, and C. Ersoy, "Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches," *IEEE Access*, vol. 8, pp. 38 146–38 163, 2020. doi: 10.1109/ACCESS.2020.2975351.
- [163] Q. Xu, T. L. Nwe, and C. Guan, "Cluster-based analysis for personalized stress evaluation using physiological signals," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 275–281, 2015. doi: 10.1109/JBHI.2014.2311044.
- [164] Y. Shi, M. H. Nguyen, P. Blitz, B. French, S. Fisk, F. De la Torre, A. Smailagic, D. P. Siewiorek, M. al'Absi, E. Ertin, *et al.*, "Personalized stress detection from physiological measurements," in *International symposium on quality of life technology*, 2010, pp. 28–29.
- [165] F. Gasparini, A. Grossi, M. Giltri, and S. Bandini, "Personalized ppg normalization based on subject heartbeat in resting state condition," *Signals*, vol. 3, no. 2, pp. 249–265, 2022.
- [166] J. Bang, T. Hur, D. Kim, T. Huynh-The, J. Lee, Y. Han, O. Banos, J.-I. Kim, and S. Lee, "Adaptive data boosting technique for robust personalized speech emotion in emotionally-imbalanced small-sample environments," *Sensors*, vol. 18, no. 11, 2018, ISSN: 1424-8220. doi: 10.3390/s18113744. [Online]. Available: <https://www.mdpi.com/1424-8220/18/11/3744>.
- [167] A. Ramponi and B. Plank, "Neural unsupervised domain adaptation in nlp—a survey," *arXiv preprint arXiv:2006.00632*, 2020.
- [168] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," *arXiv preprint arXiv:1702.05374*, 2017.
- [169] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
- [170] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.
- [171] L. R. Goldberg, "An alternative" description of personality": The big-five factor structure.," *Journal of personality and social psychology*, vol. 59, no. 6, p. 1216, 1990.
- [172] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [173] C. P. Chen and Z. Liu, "Broad learning system: An effective and efficient incremental learning system without the need for deep architecture," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 1, pp. 10–24, 2017.

- [174] J. Kocoń, M. Gruza, J. Bielaniewicz, D. Grimling, K. Kanclerz, P. Miłkowski, and P. Kazienko, “Learning personal human biases and representations for subjective tasks in natural language processing,” in *2021 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2021, pp. 1168–1173.
- [175] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.
- [176] S. Woźniak, B. Koptyra, A. Janz, P. Kazienko, and J. Kocoń, “Personalized large language models,” *arXiv preprint arXiv:2402.09269*, 2024.
- [177] D. Fanelli, “Do pressures to publish increase scientists’ bias? an empirical support from us states data,” *PLoS one*, vol. 5, no. 4, e10271, 2010.
- [178] A. Franco, N. Malhotra, and G. Simonovits, “Publication bias in the social sciences: Unlocking the file drawer,” *Science*, vol. 345, no. 6203, pp. 1502–1505, 2014.
- [179] B. Brembs, “Prestigious science journals struggle to reach even average reliability,” *Frontiers in human neuroscience*, vol. 12, p. 327 726, 2018.
- [180] S. Saganowski, J. Komoszyńska, M. Behnke, B. Perz, Ł. D. Kaczmarek, and P. Kazienko, “Emognition Wearable Dataset 2020,” version DRAFT VERSION, *Harvard Dataverse*, 2021. doi: <https://doi.org/10.7910/DVN/R9WAF4>. [Online]. Available: <https://doi.org/10.7910/DVN/R9WAF4>.
- [181] J. T. Richardson, “Eta squared and partial eta squared as measures of effect size in educational research,” *Educational research review*, vol. 6, no. 2, pp. 135–147, 2011.
- [182] D. Lakens, “Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and anovas,” *Frontiers in psychology*, vol. 4, p. 863, 2013.
- [183] J. Cohen, *Statistical power analysis for the social sciences*. Hillsdale, NJ: Erlbaum, 1988.
- [184] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, “Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers,” *Cognition and emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.
- [185] M. Behnke, J. J. Gross, and L. D. Kaczmarek, “The role of emotions in esports performance.,” *Emotion*, 2020.
- [186] A. Marchewka, Ł. Żurawski, K. Jednoróg, and A. Grabowska, “The nencki affective picture system (naps): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database,” *Behavior research methods*, vol. 46, no. 2, pp. 596–610, 2014.

- [187] P. Jemioło, D. Storman, M. Mamica, M. Szymkowski, W. Żabicka, M. Wojtaszek-Główka, and A. Ligęza, “Datasets for automated affect and emotion recognition from cardiovascular signals using artificial intelligence—a systematic review,” *Sensors*, vol. 22, no. 7, p. 2538, 2022.
- [188] B. Perz, “Personalization of emotion recognition for everyday life using physiological signals from wearables,” in *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, IEEE, 2022, pp. 1–5.
- [189] 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.
- [190] 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, Í. 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. DOI: 10.1038/s41592-019-0686-2.
- [191] 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.
- [192] 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.
- [193] 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>.
- [194] 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.

- [195] J. P. Shaffer, “Modified sequentially rejective multiple test procedures,” *Journal of the American Statistical Association*, vol. 81, no. 395, pp. 826–831, 1986.
- [196] S. Saganowski, J. Miszczyk, D. Kunc, D. Lisouski, and P. Kazienko, “Lessons learned from developing emotion recognition system for everyday life,” in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 1047–1054.
- [197] E. Topolewska, E. Skimina, W. Strus, J. Ciecuch, and T. Rowiński, “Krótki kwestionariusz do pomiaru wielkiej piątki ipip-bfm-20,” *Roczniki Psychologiczne*, vol. 17, no. 2, pp. 367–384, 2014.
- [198] Z. Juczyński, *Narzędzia pomiaru w promocji i psychologii zdrowia. skala satysfakcji z życia*, 2001.
- [199] E. Diener, D. Wirtz, W. Tov, C. Kim-Prieto, D.-w. Choi, S. Oishi, and R. Biswas-Diener, “New well-being measures: Short scales to assess flourishing and positive and negative feelings,” *Social indicators research*, vol. 97, pp. 143–156, 2010.
- [200] B. Löwe, I. Wahl, M. Rose, C. Spitzer, H. Glaesmer, K. Wingenfeld, A. Schneider, and E. Brähler, “A 4-item measure of depression and anxiety: Validation and standardization of the patient health questionnaire-4 (phq-4) in the general population,” *Journal of affective disorders*, vol. 122, no. 1-2, pp. 86–95, 2010.
- [201] Z. Juczyński and N. Ogińska-Bulik, *Narzędzia pomiaru stresu i radzenia sobie ze stresem*. Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego, 2012.
- [202] A. C. Schat, E. K. Kelloway, and S. Desmarais, “The physical health questionnaire (phq): Construct validation of a self-report scale of somatic symptoms,” *Journal of occupational health psychology*, vol. 10, no. 4, p. 363, 2005.
- [203] H. Medland, K. De France, T. Hollenstein, D. Mussoff, and P. Koval, “Regulating emotion systems in everyday life,” *European Journal of Psychological Assessment*, 2020.
- [204] D. Preece, R. Becerra, K. Robinson, J. Dandy, and A. Allan, “The psychometric assessment of alexithymia: Development and validation of the perth alexithymia questionnaire,” *Personality and Individual Differences*, vol. 132, pp. 32–44, 2018.
- [205] F. Larradet, R. Niewiadomski, G. Barresi, and L. S. Mattos, “Appraisal theory-based mobile app for physiological data collection and labelling in the wild,” in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 2019, pp. 752–756.

- [206] K. Hoemann, Z. Khan, M. J. Feldman, C. Nielson, M. Devlin, J. Dy, L. F. Barrett, J. B. Wormwood, and K. S. Quigley, “Context-aware experience sampling reveals the scale of variation in affective experience,” *Scientific reports*, vol. 10, no. 1, p. 12459, 2020.
- [207] K. Kroenke, R. L. Spitzer, and J. B. Williams, “The phq-9: Validity of a brief depression severity measure,” *Journal of general internal medicine*, vol. 16, no. 9, pp. 606–613, 2001.
- [208] R. R. Bouckaert and E. Frank, “Evaluating the replicability of significance tests for comparing learning algorithms,” in *Pacific-Asia conference on knowledge discovery and data mining*, Springer, 2004, pp. 3–12.
- [209] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scandinavian journal of statistics*, pp. 65–70, 1979.
- [210] N. A. Coles, J. K. Hamlin, L. L. Sullivan, T. H. Parker, and D. Altschul, “Build up big-team science,” *Nature*, vol. 601, no. 7894, pp. 505–507, 2022.
- [211] N. A. Coles, L. M. DeBruine, F. Azevedo, H. A. Baumgartner, and M. C. Frank, “‘big team’ science challenges us to reconsider authorship,” *Nature Human Behaviour*, pp. 1–3, 2023.
- [212] 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.
- [213] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, and A. Smola, “Autogluon-tabular: Robust and accurate automl for structured data,” *arXiv preprint arXiv:2003.06505*, 2020.
- [214] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, “Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting,” in *International conference on machine learning*, PMLR, 2022, pp. 27268–27286.
- [215] A. Damasio and G. B. Carvalho, “The nature of feelings: Evolutionary and neurobiological origins,” *Nature reviews neuroscience*, vol. 14, no. 2, pp. 143–152, 2013.
- [216] W. B. Cannon, “Again the james-lange and the thalamic theories of emotion.,” *Psychological Review*, vol. 38, no. 4, p. 281, 1931.
- [217] T. A. D’Amelio, N. M. Bruno, L. A. Bugnon, F. Zamberlan, and E. Tagliacuzzi, “Affective computing as a tool for understanding emotion dynamics from physiology: A predictive modeling study of arousal and valence,” in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, IEEE, 2023, pp. 1–7.
- [218] M. C. Frank, M. Braginsky, J. Cachia, N. Coles, T. Hardwicke, R. Hawkins, M. B. Mathur, and R. Williams, *Experimentology: An open science approach to experimental psychology methods*, 2023.

- [219] T. Yarkoni, “The generalizability crisis,” *Behavioral and Brain Sciences*, vol. 45, e1, 2022.
- [220] A. Almaatouq, T. L. Griffiths, J. W. Suchow, M. E. Whiting, J. Evans, and D. J. Watts, “Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences,” *Behavioral and Brain Sciences*, vol. 47, e33, 2024.
- [221] J. L. Tracy and D. Randles, “Four models of basic emotions: A review of ekman and cordaro, izard, levenson, and panksepp and watt,” *Emotion review*, vol. 3, no. 4, pp. 397–405, 2011.
- [222] A. H. Sham, K. Aktas, D. Rizhinashvili, D. Kuklianov, F. Alisinanoglu, I. Ofodile, C. Ozcinar, and G. Anbarjafari, “Ethical ai in facial expression analysis: Racial bias,” *Signal, Image and Video Processing*, vol. 17, no. 2, pp. 399–406, 2023.
- [223] B. Haibe-Kains, G. A. Adam, A. Hosny, F. Khodakarami, M. A. Q. C. (S. B. of Directors Shraddha Thakkar 35 Kusko Rebecca 36 Sansone Susanna-Assunta 37 Tong Weida 35 Wolfinger Russ D. 38 Mason Christopher E. 39 Jones Wendell 40 Dopazo Joaquin 41 Furlanello Cesare 42, L. Waldron, B. Wang, C. McIntosh, A. Goldenberg, A. Kundaje, *et al.*, “Transparency and reproducibility in artificial intelligence,” *Nature*, vol. 586, no. 7829, E14–E16, 2020.
- [224] P. Obels, D. Lakens, N. A. Coles, J. Gottfried, and S. A. Green, “Analysis of open data and computational reproducibility in registered reports in psychology,” *Advances in Methods and Practices in Psychological Science*, vol. 3, no. 2, pp. 229–237, 2020.
- [225] S. Lohr, “Universities and tech giants back national cloud computing project,” *The New York Times*, 2020.
- [226] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, “Amigos: A dataset for affect, personality and mood research on individuals and groups,” *IEEE transactions on affective computing*, vol. 12, no. 2, pp. 479–493, 2018.
- [227] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, “Ascertain: Emotion and personality recognition using commercial sensors,” *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2016.
- [228] S. Katsigiannis and N. Ramzan, “Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices,” *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 98–107, 2017.
- [229] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, “NeuroKit2: A python toolbox for neurophysiological signal processing,” *Behavior Research Methods*, vol. 53, no. 4,

- pp. 1689–1696, Feb. 2021. doi: 10.3758/s13428-020-01516-y. [Online]. Available: <https://doi.org/10.3758/s13428-020-01516-y>.
- [230] A. Gu, K. Goel, and C. Ré, “Efficiently modeling long sequences with structured state spaces,” *arXiv preprint arXiv:2111.00396*, 2021.
- [231] K. Mundnich, B. M. Booth, M. l’Hommedieu, T. Feng, B. Girault, J. L’hommedieu, M. Wildman, S. Skaaden, A. Nadarajan, J. L. Villatte, *et al.*, “Tiles-2018, a longitudinal physiologic and behavioral data set of hospital workers,” *Scientific Data*, vol. 7, no. 1, p. 354, 2020.
- [232] M. Hollander and D. Wolfe, *Nonparametric statistical methods*. New York: John Wiley & Sons, 1973.
- [233] W. J. Conover and R. L. Iman, “Rank transformations as a bridge between parametric and nonparametric statistics,” *The American Statistician*, vol. 35, no. 3, pp. 124–129, 1981.
- [234] W. J. Conover and R. L. Iman, “Multiple-comparisons procedures. informal report,” Los Alamos National Lab.(LANL), Los Alamos, NM (United States), Tech. Rep., 1979.
- [235] W. J. Conover, *Practical nonparametric statistics*. john wiley & sons, 1999, vol. 350.
- [236] 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, pp. 258–267, 2007, ISSN: 10155759.
- [237] M. M. Bradley and P. J. Lang, “Measuring emotion: The self-assessment manikin and the semantic differential,” *J Behav Ther Exp Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.
- [238] K. Oatley and P. N. Johnson-Laird, “Cognitive approaches to emotions,” *Trends Cogn. Sci.*, vol. 18, no. 3, pp. 134–140, 2014.
- [239] J. D. Morris, “Observations: Sam: The self-assessment manikin an efficient cross-cultural measurement of emotional response 1,” *Journal of advertising research*, vol. 35, no. 6, pp. 63–68, 1995.
- [240] P. Schmidt, A. Reiss, R. Dürichen, and K. Van Laerhoven, “Labelling affective states" in the wild" practical guidelines and lessons learned,” in *Proc. 2018 ACM Int. Joint Conf. Symp. Perv. Ubiq. Comp. Wear. Comp.*, 2018, pp. 654–659.
- [241] G. Matthews, S. E. Campbell, S. Falconer, L. A. Joyner, J. Huggins, K. Gilliland, R. Grier, and J. S. Warm, “Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry,” *Emotion*, vol. 2, no. 4, p. 315, 2002.

- [242] M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas, “The mini-ipip scales: Tiny-yet-effective measures of the big five factors of personality.” *Psychological assessment*, vol. 18, no. 2, p. 192, 2006.
- [243] E. Topolewska, E. Skimina, W. Strus, J. Ciecuch, and T. Rowiński, “Krótki kwestionariusz do pomiaru wielkiej piątki ipip-bfm-20,” *Roczniki Psychologiczne*, vol. 17, no. 2, pp. 367–384, 2019.
- [244] C. E. Carney, D. J. Buysse, S. Ancoli-Israel, J. D. Edinger, A. D. Krystal, K. L. Lichstein, and C. M. Morin, “The consensus sleep diary: Standardizing prospective sleep self-monitoring,” *Sleep*, vol. 35, no. 2, pp. 287–302, 2012.
- [245] A. M. Gordon and W. B. Mendes, “A large-scale study of stress, emotions, and blood pressure in daily life using a digital platform,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 31, e2105573118, 2021.
- [246] P. Gertler and R. L. Tate, “Are single item mood scales (sims) valid for people with traumatic brain injury?” *Brain Injury*, vol. 34, no. 5, pp. 653–664, 2020.
- [247] J. O. Pinzon-Arenas, L. Mercado-Diaz, J. Tejada, F. Marmolejo-Ramos, C. Barrera-Causil, J. I. Padilla, R. Ospina, and H. Posada-Quintero, “Deep learning analysis of electrophysiological series for continuous emotional state detection,” in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, IEEE, 2023, pp. 1–8.

Appendix A

CRITICAL LITERATURE REVIEW

Table A.1: Emotional models, ground truth, and machine learning (ML) problems (from [16]). Env. – Environment; Lab – lab study; Field – field study; FC – field with constraints; 'SA' – self-assessment; '?' means it was not described but deduced by us only; '/' separates different setups considered; n -class denotes a multiclass problem solved

Year	Paper	Scenario; Env.	Initial emotional model	Discrete emotions + neutral	Dimensions	Modification of the initial model	Ground truth	ML problem
2002	Nasoz et al. [87]	1?/3?; Lab	Own	4+1: anger, fear, sadness, frustration, neutral	–	–	SA	5-class
2004, 2006	Lisetti et al. [89, 90]	2; Lab	Own	6: sad, anger, fear, surprise, frustration, amusement	–	–	labels assigned to stimuli?	6-class
2010, 2011	Rattanyu et al. [95, 96]	3; Lab	Plutchik & Circumplex [45]	5+1: anger, fear, disgust, sadness, joy + neutral \approx Ekman-Friesen [39] w/o surprise	–	selection + neutral	SA if agreed with picture label	6-class?
2015	Guo et al. [97]	3; Lab	Own	4+1: anger, fear, sadness, happiness, peace	–	–	video labels confirmed by SA?	5-class
2016	Guo et al. [98]	3; Lab	Own	4+1: angry, fear, sad, happy, relax	–	Also grouped into positive-negative	stimuli (video) labels confirmed by SA?	binary / 5 x binary (one against all)
2016	Exler et al. [76]	7; Field	3D MDMQ [236]	–	1: valence (3 levels)	selection from 3D	SA	3-class
2017	Pollreisz and TaheriNejad [78]	1; Lab	Own	4: happy, sad, angry, pain	–	–	SA	4-class
2017, 2018	Kanjo et al. [103, 104]	6; FC	SAM [237]	–	1: valence (5 levels)	–	SA	5-class
2017	Kim et al. [105]	6; FC	PAD-SAM	–	3: pleasure, arousal, dominance (9 levels each)	–	SA	regression
2017	Nguyen et al. [106]	7; Field	Oatley-Johnson [238]	5+1: fear, anger, sadness, disgust, happiness + neutral \approx Ekman-Friesen [39] w/o surprise	–	grouped into: negative-neutral-positive	SA	3-class?

2018	Quiroz et al. [99]	3; Lab	Own	2: happy, sad / 2+1: happy, sad, neutral	–	–	stimuli labels + SA (PANAS)	binary / 3-class
2018	Feng et al. [102]	5; Lab	Circumplex [46]	2+1: joy, boredom, acceptance=neutral	–	selection	external experts	3 x binary (one vs. another) / 3-class
2018	Zhao et al. [79]	1; Lab	Dimensional	–	2: arousal (low, high), valence (low, high)	equivalent to quadrants	SA	4-class / 2 x binary
2018	Albraikan et al. [80]	1/2; Lab	Dimensional	4+1: cheer, sadness, erotic, horror, neutral	2: arousal (calm, medium, activated), valence (unpleasant, neutral, pleasant)	9 keywords located in the 2D space (3 valence x 3 arousal)	stimuli (video) labels / SA / keywords	5-class / 2 x 3-class
2018	Albraikan et al. [101]	4; Lab	Own	4+1: happy, sad, love, fear, neutral	–	selection of 3 emotions	trained own system vs. SA	3-class / 5-class
2018	Nakisa et al. [81]	1; Lab	3D (SAM-like)	–	2: arousal, valence	keywords located in quadrants	SA (keywords)	4-class
2018	Dao et al. [63]	6; FC	?	6?: excited, bored, stressed, relaxed, happy, serene	–	–	system / SA	6-class?
2018	Setiawan et al. [108]	?; Lab	Own	4: joy, sad, stress, calm	2: arousal, valence (2-valued each?)	–	?	2 x binary?
2018	Hu et al. [88]	1?/5?; Lab?	Krech et al. [120]	4+1: happy, angry, fear, sad, normal	–	grief from [120] was replaced with sad	SA? or expert?	5-class
2018	Ragot et al. [82]	1; Lab	SAM [237]	–	2: arousal, valence	selection= quadrants	SA	2 x binary
2019	Bulagang et al. [91]	2; Lab	Circumplex [46] - dimensional	–	2: arousal, valence	quadrants	stimuli labels	4-class?
2019	Romeo et al. [83]	1; Lab	SAM [237] + liking	–	2: arousal, valence (9 levels each)	dimension selection, level reduction to low-high	SA	2 x binary

2019	Wampfler et al. [84]	1; Lab	SAM [237]	3/5/6 (affective regions)	2: arousal, valence	region identification	regions based on SA	3/5/6-class / clustering in arousal-valence space
2019	Schmidt et al. [11]	7; Field	SAM [239]	–	2: arousal, valence (3 levels each)	dim selection	SA (EMA [240])	2 x 3-class / multi-task with STAI & stress
2020	Shu et al. [92]	2; Lab	Dimensional	–	1: valence (3 levels)	–	stimuli (video) labels	3-class
2020	Setiawan et al. [93]	2; Lab	Own	4: happiness, angry, sad, calm / 7 for facial expressions	–	–	stimuli labels	4-class?
2020	Martens et al. [85]	1; Lab	Own	–	5: interest, energy, valence, focus, tension	–	SA: DSSQ [241]	5 x regression
2020	Kadoya et al. [107]	7?; Field	Circumplex [46]	4+1: happy, angry, relaxed, sad, neutral	–	selection + neutral	SA?	multiclass?
2020	Tizzano et al. [100]	3; Lab	Own	2: happy, sad / 2+1: happy, sad, neutral	–	–	stimuli labels + SA (PANAS)	binary / 3-class
2020	Majumder et al. [94]	2?; Lab	Own	3+1: happy, angry, sad, neutral	–	–	stimuli labels?	4-class
2020	Saxena et al. [86]	1; Lab	Own	5+1: amusement, anger, disgust, fear, sad, neutral / 2+1: positive, negative, neutral / 5+1+anxiety / 2+1+anxiety	–	–	SA	3/6-class / regression

Table A.2: Main differences in emotion recognition between lab study and field study (from [28]). ' + ' denotes an advantage; ' - ' is a disadvantage; ' ± ' means an aspect has both, positive and negative sides

Category	Lab study	Field study
Emotions experienced	<ul style="list-style-type: none"> - In controlled environment - Impacted by unnatural conditions - Limited to the prepared stimuli + Beginning and end determined by the stimuli 	<ul style="list-style-type: none"> + In natural context + Full range of emotions - Occurrence is difficult to capture - Hard to determine the beginning and end
Stimuli	<ul style="list-style-type: none"> ± Planned and prepared, e.g., videos, images, music, tasks + Fully controlled by researchers, may be interrupted + May be annotated + Known duration + No distractions nor unexpected stimuli + Condensed sequence of stimulants separated by wash out 	<ul style="list-style-type: none"> + Daily life stimuli - Unknown stimuli - No stimuli label - No starting point - Unknown duration - Out of researcher's control - Susceptible to life conditions, e.g., drugs, fatigue
Labeling (ground truth)	<ul style="list-style-type: none"> + Self-assessment + Expert-annotated stimuli + Observed and derived by external experts 	<ul style="list-style-type: none"> - Mainly self-assessment + Nearby person (relative, friend)
Self-assessment	<ul style="list-style-type: none"> + Detailed + Often + Trigger time easy to determine + Triggered and filled out right after each stimuli 	<ul style="list-style-type: none"> - Limited scope - Sporadic - Triggering time is difficult to determine ± Self-, event-, activity-, randomly-triggered, schedule, reasoning - Usually delayed participant's response [76]
Measuring physiology / devices	<ul style="list-style-type: none"> + Medical-level, precise devices + Devices can be large and wired + Many devices simultaneously possible + External devices possible, e.g., multiple cameras + No battery problem - Stressful condition + High-quality signal / data (little external interference) + Stationary position (usually sitting) 	<ul style="list-style-type: none"> - Lower quality of sensors and signals [77] + Personal, convenient, useful wearables - Only few devices feasible ± Battery-efficient wearables + Convenient and unnoticeable measuring - Artifacts caused by the movement and field conditions ± Data transfer to server (in real-time / post-session) - Lack of data when wearable is off / not worn - 24/7 technical support required
Additional factors	<ul style="list-style-type: none"> + Static environment (temperature, lighting, etc.) + Meta-questions (e.g., health issues, time past since last coffee/activity/sleep) ± Relatively small amount of data 	<ul style="list-style-type: none"> - Variable environment - No meta-question ± Large amount of data to be collected and processed

Table A.3: Machine learning-related techniques and measures applied in SLR studies (from [28]). '?' means it was not clarified but concluded by us only; '/' separates different setups considered; n -class denotes a multiclass problem solved

Year	Paper	Classification type	ML models applied	Tested hyperparameters	Quality measures	Imbalance in learning samples	Statistical tests on results
2002	Nasoz et al. [87]	5-class	KNN, LDA	no info	accuracy	balanced?	no
2004, 2006	Lisetti et al. [89, 90]	6-class	KNN, LDA, MBP	no info	accuracy	balanced	no
2010, 2011	Rattanyu et al. [95, 96]	6-class?	adaptable KNN, LDA	no info	accuracy	no info	ANOVA, LSD
2015	Guo et al. [97]	5-class	DT based on their own rules	none	accuracy	balanced	no
2016	Exler et al. [76]	3-class	DT	no info	accuracy	not considered; high imbalance	no
2016	Guo et al. [98]	binary / 5 x binary (one against all)	SVM	various kernels	accuracy	binary: small imbalance / 5 x binary: equal size sampling	no
2017	Pollreisz and TaheriNejad [78]	4-class	DT based on their own rules	no info	accuracy, conf. level (probability)	not considered; small imbalance	no
2017	Kanjo et al. [103]	5-class	KNN, RF, stacking with NB as a learner, SVM	no info	accuracy, F-measure	no info	no
2017	Kim et al. [105]	regression	mixed linear model	no info	-	no info	p-value, analysis of beta coeff.
2017	Nguyen et al. [106]	3-class?	DT, KNN, SVM	none	accuracy	balanced	no
2018	Quiroz et al. [99]	binary / 3-class	RF, LR	none	accuracy, F-measure, AUC	balanced	p-value
2018	Feng et al. [102]	3 x binary (one vs. another) / 3-class	KNN, SVM	SVM: 3 kernels; KNN: $k = 1, 3, 5$	accuracy, AUC, precision, recall	not considered; imbalanced	no
2018	Zhao et al. [79]	4-class / 2 x binary	NB, NN, RF, SVM	no info	correct classification ratio	no info	no
2018	Albraikan et al. [80]	5-class / 2 x 3-class	no info	no info	accuracy, F-measure	their dataset balanced? / MAHNOB imbalanced	no info

Year	Paper	Classification type	ML models applied	Tested hyperparameters	Quality measures	Imbalance in learning samples	Statistical tests on results
2018	Albraikan et al. [101]	3-class / 5-class	no info	no info	accuracy, statistical test for difference	no info	McNemar's test
2018	Nakisa et al. [81]	4-class	LSTM, MLP, SVM	optimization algorithms (DE, PSO, SA, RS, TPE), batch size, no. of hidden neurons	accuracy	small imbalance?	ANOVA
2018	Dao et al. [63]	6-class?	no info	no info	no info	no info	no
2018	Setiawan et al. [108]	2 x binary?	DT, LR, RF, SVM	no info	accuracy	no info	no
2018	Kanjo et al. [104]	5-class	CNN, CNN-LSTM, MLP	no info	accuracy, F-measure, precision, recall, error rate, RMSE, confusion matrix	no info	no
2018	Hu et al. [88]	5-class	SVM	no info	accuracy	not considered; high imbalance	no
2018	Ragot et al. [82]	2 x binary	SVM	no info	accuracy	no info	no
2019	Bulagang et al. [91]	4-class?	SVM	gamma values?	accuracy	balanced?	no
2019	Romeo et al. [83]	2 x binary	NB, RF, various MILs, various SVMs	RF: no. of trees; mil-Boost: weak learners	accuracy, macro F-measure, ROC curves, confusion matrices	not considered; imbalanced	Wilcoxon's signed-rank
2019	Wampfler et al. [84]	3/5/6-class / clustering in arousal-valence space	Gaussian NB, KM, KNN, RF, SVM	randomized search with 100 iterations (no details)	accuracy, micro-avg AUC, macro-avg AUC	high imbalance; RF with balanced class weights, macro-avg AUC	no
2019	Schmidt et al. [11]	2 x 3-class / multi-task with STAI & stress	CAE, CNN, DT, randomized DT, RF	Adam optimization	macro F-measure	high imbalance; converting Likert scales into bins, but data still imbalanced	Pearson correlation coeff.

Year	Paper	Classification type	ML models applied	Tested hyperparameters	Quality measures	Imbalance in learning samples	Statistical tests on results
2020	Shu et al. [92]	3-class	AdaBoost, DT, Gradient Boosting DT, KNN, RF	no info	accuracy	balanced	no
2020	Setiawan et al. [93]	4-class?	DF, DT, KNN, NB, RF, SVM	no info	accuracy	balanced?	no
2020	Martens et al. [85]	5 x regression	gradient boosting, Ridge Regression	no info	MAE, RMSE, Pearson correlation coefficient	no info	student's t-tests
2020	Kadoya et al. [107]	multiclass?	no info	no info	no info	not considered; high imbalance	no
2020	Tizzano et al. [100]	binary / 3-class	SVM, GMM, LSTM	Adam optimization	accuracy	balanced	no
2020	Majumder et al. [94]	4-class	TreeBagger Bootstrap	none	accuracy, confusion matrix	balanced?	no
2020	Saxena et al. [86]	3/6-class / regression	Gaussian SVM, Cubic KNN, Weighted KNN, Ensemble Bagged Trees, Ensemble Boosted Trees, NN	no info	accuracy, confusion matrices	balanced	no

Appendix B

IN-THE-LABORATORY DATA COLLECTION


Proszę wskazać, w jakim stopniu odczuwasz emocje.

	wcale lub nieznacznie	trochę	średnio	bardzo	wyjątkowo mocno
zachwyt	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
obrzydzenie	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
zaskoczenie	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
złość	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
radość	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
pragnienie	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
strach	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
rozbawienie	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
smutek	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

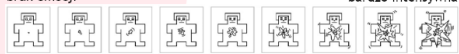
(a) Questionnaire for discrete emotions.

Proszę wskazać, które obrazki najlepiej przedstawiają Twoje odczucia.


nieprzyjemność przyjemność



brak emocji bardzo intensywna



odpycha/unikać przyciąga/dążyć



(b) Questionnaire for valence, arousal, and motivation.

Figure B.1: The original (Polish) version of self-reports used in the Emognition study [24].

Table B.1: Results of Repeated Measures Analysis of Variance for Differences Between Conditions in Self-reported Emotions (from [24]).

Self-reports	Amusement		Anger		Awe		Disgust		Enthusiasm		Fear		Liking		Sadness		Surprise		Baseline		Neutral		rm ANOVA		η^2_p
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	F	df	
Amusement	3.37	0.98	1.16	0.43	1.3	0.64	2.33	1.11	1.47	0.85	1.3	0.6	1.6	0.85	1.07	0.26	1.26	0.49	1.37	0.72	1.14	0.41	45.37***	5.32, 223.57	0.52
Anger	1.09	0.37	2.7	1.26	1.14	0.41	1.19	0.39	0	1.28	0.59	1.02	0.15	1.47	0.67	1.02	0.15	1.07	0.34	1.09	0.37	38.21***	2.96, 124.23	0.48	
Awe	1.42	0.73	1.16	0.57	2.86	1.08	1.14	0.47	2.74	1.18	1.12	0.39	2.74	1.07	1.12	0.5	1.02	0.15	1.19	0.5	1.4	0.62	49.04***	4.80, 201.76	0.54
Disgust	1.4	0.93	2.84	1.25	1	0	3.49	1.14	1.02	0.15	1.47	0.7	1.05	0.21	1.26	0.58	1.14	0.41	1.07	0.34	0	70.17***	3.97, 166.75	0.63	
Enthusiasm	2.3	0.99	1.02	0.15	2.37	1.2	1.35	0.78	2.95	1.05	1.02	0.15	2.58	1.22	1.02	0.15	1.02	0.15	1.35	0.69	1.35	0.61	46.60***	4.06, 170.42	0.53
Fear	1.07	0.34	2.19	1.07	1.05	0.3	1.26	0.49	1	0	2.7	1.17	1	0	1.26	0.54	1.6	0.9	1.07	0.34	1.21	0.47	37.57***	4.09, 171.88	0.47
Liking	1.26	0.58	1	0	2.44	1.22	1	0	2.14	1.1	1.05	0.21	3.16	1.19	1.02	0.15	1	0	1.07	0.34	1.35	0.65	56.76***	3.64, 152.92	0.57
Sadness	1.05	0.21	2.65	1.11	1.42	0.73	1.19	0.39	1.02	0.15	1.44	0.77	1.02	0.15	3.16	1.02	1.12	0.39	1.09	0.37	1.23	0.43	68.50***	4.30, 180.45	0.62
Surprise	2.58	1.1	2.49	1.24	1.42	0.7	2.56	1.2	1.33	0.64	2.14	1.19	1.26	0.49	1.42	0.63	3.33	1.02	1.21	0.47	1.51	0.7	36.40***	6.64, 279.08	0.46
Valence	6.58	1.43	2.6	1.35	6.47	1.42	3.91	2.09	6.98	1.24	3.65	1.54	7.3	1.28	3.37	0.9	4.86	1.04	5.02	0.94	5.35	0.78	64.25***	5.31, 222.87	0.6
Arousal	4.95	1.54	5.79	2.16	4.05	2.03	5.33	1.82	4.56	1.99	4.51	2.13	4.6	1.93	4.44	2.07	4.07	2.11	1.98	1.55	2.37	1.33	25.57***	6.50, 272.93	0.38
Motivation	5.84	1.65	2.65	1.6	7.07	1.56	2.4	1.72	7.23	1.39	3.02	1.67	7.56	1.28	3.6	1.43	4.79	1.34	4.42	1.22	5.51	1.18	78.69***	6.99, 293.55	0.65

Note. The significant results of repeated measures analysis of variance indicates differences in self-reported emotions between film clip conditions (e.g., differences in self-reported amusement between amusing film clip and sad film clip, angry film clip etc.). M = Mean, SD = Standard Deviation, F = F-Ratio calculated by dividing the mean squares for the variable by its error mean squares, ***p .001.

Table B.2: Results of Repeated Measures Analysis of Variance for Differences Within Conditions in Self-reported Emotions (from [24]).

Film Clips	Amusement		Anger		Awe		Disgust		Enthusiasm		Fear		Liking		Sadness		Surprise		Baseline		Neutral		rm ANOVA		η^2_p
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	F	df	
Amusement	3.37	0.98	1.09	0.37	1.42	0.73	1.4	0.93	2.3	0.99	1.07	0.34	1.26	0.58	1.05	0.21	2.58	1.1	57.63***	3.69, 154.91	0.58				0.58
Anger	1.16	0.43	2.7	1.26	1.16	0.57	2.84	1.25	1.02	0.15	2.19	1.07	1	0	2.65	1.11	2.49	1.24	38.26***	4.52, 189.84	0.48				0.48
Awe	1.3	0.64	1.14	0.41	2.86	1.08	1	0	2.37	1.2	1.05	0.3	2.44	1.22	1.42	0.73	1.42	0.7	39.04***	3.62, 152.09	0.48				0.48
Disgust	2.33	1.11	1.19	0.39	1.14	0.47	3.49	1.14	1.35	0.78	1.26	0.49	1	0	1.19	0.39	2.56	1.2	58.73***	3.50, 146.79	0.58				0.58
Enthusiasm	1.47	0.85	1	0	2.74	1.18	1.02	0.15	2.95	1.05	1	0	2.14	1.1	1.02	0.15	1.33	0.64	64.86***	3.08, 129.23	0.61				0.61
Fear	1.3	0.6	1.28	0.59	1.12	0.39	1.47	0.7	1.02	0.15	2.7	1.17	1.05	0.21	1.44	0.77	2.14	1.19	29.08***	3.50, 146.95	0.41				0.41
Liking	1.6	0.85	1.02	0.15	2.74	1.07	1.05	0.21	2.58	1.22	1	0	3.16	1.19	1.02	0.15	1.26	0.49	71.87***	2.64, 110.79	0.63				0.63
Sadness	1.07	0.26	1.47	0.67	1.12	0.5	1.26	0.58	1.02	0.15	1.26	0.54	1.02	0.15	3.16	1.02	1.42	0.63	66.77***	3.83, 160.79	0.62				0.62
Surprise	1.26	0.49	1.02	0.15	1.02	0.15	1.14	0.41	1.02	0.15	1.6	0.9	1	0	1.12	0.39	3.33	1.02	94.21***	2.59, 108.926	0.69				0.69
Baseline	1.37	0.72	1.07	0.34	1.19	0.5	1.07	0.34	1.35	0.69	1.07	0.34	1.07	0.34	1.09	0.37	1.21	0.47	5.33***	2.53, 106.07	0.11				0.11
Neutral	1.14	0.41	1.09	0.37	1.4	0.62	1	0	1.35	0.61	1.21	0.47	1.35	0.65	1.23	0.43	1.51	0.7	5.34***	4.99, 209.62	0.11				0.11

Note. The significant results of repeated measures analysis of variance indicates differences in self-reported emotions within film clip conditions (e.g., differences in amusing film clip between self-reported amusement and sadness, anger etc.). M = Mean, SD = Standard Deviation, F = F-Ratio calculated by dividing the mean squares for the variable by its error mean squares, ***p .001.

*Appendix C***EMOTIONS IN THE WILD****C.1 Questionnaires utilized for experiments**

Although we utilized many more questionnaires to collect contextual data in the LarField study, we were not able to utilize all of them in our experiments. Below, we present the Polish versions that we used for research in this dissertation, with references to their original English versions if we used such. Participants filled out personality and demography questionnaires at the start of the study. Daily questionnaires were shown to them in the app – morning and evening questionnaires once a day, and emotion questionnaires six times a day or more if triggered manually. In the Emognition app, participants answered daily questionnaires using sliders or checkboxes, depending on the specific questions¹. Polish versions of questionnaires were created by members of the Emognition research group unless otherwise noted.

C.1.1 Personality and demography questionnaires**Personality questionnaire**

¹The outline of the Emognition application can be seen in the demonstration video: <https://www.youtube.com/watch?v=qk3DFmRKKlw>

Osobowość: Kwestionariusz IPIP-BFM-20

Przeczytaj uważnie poniższe zdania, opisujące różne zachowania, uczucia i myśli ludzi. Zastanów się nad każdym z nich – w jakim stopniu opisuje ono również Ciebie takiego/taką, jakim/jaką zwykle jesteś? Ludzie są bardzo różni, więc nie ma tu dobrych ani złych odpowiedzi. Za każdym razem po prostu szczerze odpowiedz na pytanie, w jakim stopniu dane stwierdzenie opisuje Ciebie.

Posługuj się następującą skalą:

- 1 – całkowicie nie trafnie mnie opisuje
- 2 – raczej nie trafnie mnie opisuje
- 3 – trochę trafnie, a trochę nie trafnie mnie opisuje
- 4 – raczej trafnie mnie opisuje
- 5 – całkowicie trafnie mnie opisuje

1. Jestem duszą towarzystwa.	1	2	3	4	5
2. Niezbyt obchodzą mnie inni ludzie.	1	2	3	4	5
3. Zostawiam moje rzeczy gdzie popadnie.	1	2	3	4	5
4. Zwykle jestem zrelaksowany/a.	1	2	3	4	5
5. Mam bogate słownictwo.	1	2	3	4	5
6. Trzymam się z boku.	1	2	3	4	5
7. Jestem wyrozumiały/a dla uczuć innych ludzi.	1	2	3	4	5
8. Bez zwłoki <u>wypełniam</u> codzienne obowiązki.	1	2	3	4	5
9. Często martwię się czymś.	1	2	3	4	5
10. Mam trudności ze zrozumieniem abstrakcyjnych pojęć.	1	2	3	4	5
11. Rozmawiam z wieloma różnymi ludźmi na przyjęciach.	1	2	3	4	5
12. Nie interesują mnie problemy innych ludzi.	1	2	3	4	5
13. Często zapominam odkładać rzeczy na miejsce.	1	2	3	4	5
14. Rzadko czuję się przygnębiony/a.	1	2	3	4	5
15. Mam głowę pełną pomysłów.	1	2	3	4	5
16. Wśród nieznajomych jestem małomówny/a.	1	2	3	4	5
17. Znajduję czas dla innych.	1	2	3	4	5
18. Postępuję zgodnie z harmonogramem.	1	2	3	4	5
19. Często miewam huśtawki nastrojów.	1	2	3	4	5
20. Nie mam zbyt bogatej wyobraźni.	1	2	3	4	5

Figure C.1: Personality questionnaire used in the Emognition study [25] – IPIP-BFM-20 [242]. Polish adaptation by Topolewska et al. [243].

Demography questionnaire

[Płeć / Gender] Określam swoją płeć jako...

- Kobieta • Mężczyzna • Niebinarny • Wolę się samookreślić (odpowiedź otwarta)
- Wolę nie mówić

[Wiek / Age] Jaki jest Pana/Pani wiek w latach (proszę wpisać tylko liczbę)?

- Odpowiedź otwarta / Free text

[Narodowość / Nationality] Jakiej jest Pani/Pana narodowości? (np. Polak)

- Odpowiedź otwarta / Free text

[Język / Language] Czy język polski jest Twoim pierwszym lub ojczystym językiem?

- Tak • Nie

[Dzieci / Children] Czy ma Pani/Pan dzieci w wieku poniżej 18 lat?

- Tak: proszę wymienić liczbę dzieci w polu odpowiedzi otwartej (Odpowiedź otwarta / Free text) • Nie

[Opieka / Caregiving] Czy ma Pani/Pan inne obowiązki związane z opieką nad dziećmi?

- Tak • Nie

[Dochody / Income] Twoja sytuacja ekonomiczna w porównaniu do przeciętnej osoby w Twoim kraju jest:

- zdecydowanie gorsza • gorsza • raczej gorsza • podobna / taka sama • raczej lepsza • lepsza • zdecydowanie lepsza

[Sektor zatrudnienia / Employment sector]

- Przetwórstwo przemysłowe • Rolnictwo, leśnictwo, łowiectwo, rybactwo • Handel; naprawa pojazdów samochodowych • Edukacja • Administracja publiczna i obrona narodowa; obowiązkowe zabezpieczenia społeczne • Budownictwo • Transport i gospodarka magazynowa • Opieka zdrowotna i pomoc społeczna • Działalność profesjonalna, naukowa i techniczna • Administrowanie i działalność wspierająca • Informacja i komunikacja • Działalność finansowa i ubezpieczeniowa • Pozostała Działalność usługowa • Zakwaterowanie i gastronomia • Obsługa rynku nieruchomości • Dostawa wody, gospodarowanie Ściekami i odpadami • Działalność związana z kulturą, rozrywką i rekreacją • Górnictwo i wydobywanie • Inny: (Odpowiedź otwarta / Free text)

[Status związku / Relationship status] Jaki jest Twój status związku?

- Samotny • W związku, ale nie mieszkam razem • Zamężna lub mieszkająca razem • Owdowiały • Rozwiedziony lub w separacji • Inne (proszę określić) (Odpowiedź otwarta / Free text)

[Wielkość gospodarstwa domowego / Number of people in a household] Ile osób, łącznie z Panem/Panią, mieszka w gospodarstwie domowym?

- Odpowiedź otwarta / Free text

[Wykształcenie / Education] Jaki jest Pana(i) poziom wykształcenia?

- Podstawowe • Średnie • Wyższe (w trakcie) • Wyższe

[Obecnie studiujesz / Are you studying] Czy obecnie studiujesz?

- Tak • Nie

[Zatrudnienie / Employment status] Jaki jest Twój status zatrudnienia?

- Pełny etat • W niepełnym wymiarze czasu pracy • Dorywczo • Samozatrudniony • Bezrobotny (i poszukujący pracy) • Bezrobotni (nie poszukujący pracy)
 - Nie pracuję zarobkowo (np. opieka, pomoc domowa, emerytura, wolontariat, student, uczeń) • Zamierzam wkrótce rozpocząć nową pracę • Inny (proszę określić)
- (Odpowiedź otwarta / Free text)

[Religia / Religious practices] Czy jest Pan(i) osobą religijną?

- Nie • Raczej nie • Raczej tak • Tak

[Poglądy polityczne / Political views] (Odpowiedź nieobowiązkowa / Answer not mandatory) Z jakimi poglądami politycznymi się utożsamiasz na osi ekonomicznej oraz światopoglądowej?

Oś ekonomiczna:

- Lewicowe (rozwinięty interwencjonizm państwowy, wysokie podatki, wyższe podatki dla bogatszych, rozwinięta polityka socjalna)
- Prawicowe (dominacja własności prywatnej, ograniczenie interwencjonizmu państwowego, niskie podatki, ograniczone ramy polityki socjalnej państwa)

Oś światopoglądowa/społeczna:

- Liberalne (akceptacja związków homoseksualnych, transseksualizmu, aborcji, egalitarność)
- Konserwatywne (ważne wartości to rodzina, naród, tradycja, hierarchia społeczna)

C.1.2 Daily questionnaires

Morning questionnaire

1. 1 - bardzo źle
7 - bardzo dobrze
2. 1 - wcale niewypoczęty/a
7 - w pełni wypoczęty/a

1. Jak oceniasz jakość swojego snu ostatniej nocy?	1	2	3	4	5	6	7
2. Jak czuleś/aś się dziś rano?	1	2	3	4	5	6	7

- 0 – zupełnie się nie zgadzam
1 – nie zgadzam się
2 – nie mam zdania
3 – zgadzam się
4 – całkowicie się zgadzam

1. Czuję się zestresowany, niespokojny, przytłoczony.	0	1	2	3	4
2. Czuję, że panuję nad sytuacją, dobrze sobie radzę, mam wszystko pod kontrolą.	0	1	2	3	4

Figure C.2: Morning sleep [244] and stress [245] questionnaire used in the Emognition study [25].

Evening questionnaire

- 0 – zupełnie się nie zgadzam
1 – nie zgadzam się
2 – nie mam zdania
3 – zgadzam się
4 – całkowicie się zgadzam

1. Czuję, że wszystko mnie teraz przytłacza.	0	1	2	3	4
2. Czuję, że wszystko jest teraz nieprzewidywalne.	0	1	2	3	4

- 0 – źle
1 – średnio
2 - dobrze
3 - bardzo dobrze
4 – doskonale

1. Jak oceniasz stan swojego zdrowia?	0	1	2	3	4
---------------------------------------	---	---	---	---	---

Jak się dzisiaj czuleś/aś?

0 - Gorzej niż kiedykolwiek	100 - Lepiej niż kiedykolwiek
-----------------------------	-------------------------------

Figure C.3: Evening health [245, 246] and stress [245] questionnaire used in the Emognition study [25].

Emotions questionnaire

Czy właśnie czujesz silną emocję? 0/1

Jak się czułaś:

0 - Negatywnie	100 - Pozytywnie
0 - Ospały	100 - bardzo pobudzony

Źródło: Saganowski, S., Behnke, M., Komoszyńska, J., Kunc, D., Perz, B., & Kazienko, P. (2021, September). A system for collecting emotionally annotated physiological signals in daily life using wearables. In 2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 1-3). IEEE.

Figure C.4: Emotion questionnaire used in the Emognition study [15, 25].

Table C.1: Results for regression of next day morning sleep quality questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.221	0.014	-	-	-
KNN	Broad context	0.208	0.015	-2.636	0.029	*
	General	0.219	0.016	-0.405	0.686	ns
	Personality	0.214	0.017	-1.608	0.222	ns
MLP	Broad context	0.254	0.020	5.233	0.000	***
	General	0.257	0.020	5.857	0.000	***
	Personality	0.256	0.021	5.194	0.000	***
Random Forest	Broad context	0.205	0.013	-5.162	0.000	***
	General	0.211	0.014	-3.330	0.007	**
	Personality	0.206	0.013	-4.732	0.000	***
SVM	Broad context	0.205	0.016	-3.643	0.003	**
	General	0.211	0.015	-2.998	0.014	*
	Personality	0.209	0.017	-3.133	0.011	*

C.2 Results

In this section, we present supplementary tables with results obtained from experiments on the LarField dataset, using four machine learning algorithms (KNN, MLP, Random Forest, SVM) and three modeling strategies: (1) models without personal context (general), (2) models with personality as context (personality), and models with personality and demography as context (broad context). All presented classification and regression scores were computed using 10-fold cross-validation repeated 10 times (10 different random seeds, same between experiments). Comparisons between results and baselines (average annotation, test set) were done using corrected paired t-test [208]. The same test was used for pairwise comparisons between modeling methods, on results averaged between four utilized models. All p-values were corrected using the Holm–Bonferroni procedure [209].

C.2.1 Prediction results

Table C.2: Results for regression of next day morning rest questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.236	0.013	-	-	-
KNN	Broad context	0.222	0.017	-2.647	0.028	*
	General	0.237	0.015	0.461	0.646	ns
	Personality	0.230	0.017	-1.060	0.584	ns
MLP	Broad context	0.269	0.023	4.400	0.000	***
	General	0.274	0.022	5.454	0.000	***
	Personality	0.273	0.019	6.291	0.000	***
Random Forest	Broad context	0.218	0.013	-5.418	0.000	***
	General	0.221	0.013	-4.562	0.000	***
	Personality	0.219	0.013	-5.032	0.000	***
SVM	Broad context	0.217	0.016	-3.870	0.001	**
	General	0.221	0.016	-3.537	0.003	**
	Personality	0.221	0.016	-3.208	0.007	**

Table C.3: Results for regression of next day morning stress questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.300	0.013	-	-	-
KNN	Broad context	0.235	0.016	-10.205	0.000	***
	General	0.235	0.016	-10.286	0.000	***
	Personality	0.229	0.019	-10.077	0.000	***
MLP	Broad context	0.274	0.021	-3.035	0.006	**
	General	0.280	0.023	-2.120	0.037	*
	Personality	0.271	0.020	-3.628	0.001	**
Random Forest	Broad context	0.242	0.015	-10.506	0.000	***
Forest	General	0.228	0.015	-13.213	0.000	***
	Personality	0.226	0.015	-13.115	0.000	***
SVM	Broad context	0.232	0.016	-10.642	0.000	***
	General	0.226	0.017	-11.032	0.000	***
	Personality	0.227	0.017	-10.726	0.000	***

Table C.4: Results for regression of next day morning composure questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.269	0.014	-	-	-
KNN	Broad context	0.226	0.019	-6.470	0.000	***
	General	0.242	0.018	-4.465	0.000	***
	Personality	0.232	0.018	-5.729	0.000	***
MLP	Broad context	0.266	0.020	-0.422	1.349	ns
	General	0.283	0.022	1.865	0.196	ns
	Personality	0.271	0.020	0.216	0.830	ns
Random Forest	Broad context	0.225	0.015	-8.281	0.000	***
	General	0.231	0.016	-7.552	0.000	***
	Personality	0.225	0.016	-7.965	0.000	***
SVM	Broad context	0.222	0.017	-7.597	0.000	***
	General	0.230	0.018	-6.022	0.000	***
	Personality	0.227	0.017	-6.483	0.000	***

Table C.5: Results for classification of next day morning sleep quality questionnaire.

Model	Personalization	F1-macro	Std	t	p	Signif.
Baseline	-	0.445	0.017	-	-	-
KNN	Broad context	0.614	0.080	6.027	0.000	***
	General	0.514	0.069	2.973	0.019	*
	Personality	0.531	0.073	3.412	0.006	**
MLP	Broad context	0.593	0.073	5.559	0.000	***
	General	0.583	0.071	5.478	0.000	***
	Personality	0.587	0.075	5.434	0.000	***
Random Forest	Broad context	0.468	0.044	1.659	0.200	ns
SVM	General	0.465	0.046	1.388	0.168	ns
	Personality	0.478	0.055	1.869	0.258	ns
SVM	Broad context	0.472	0.046	1.839	0.207	ns
	General	0.445	0.017	0.000	0.000	***
	Personality	0.445	0.017	0.000	0.000	***

Table C.6: Results for classification of next day morning rest questionnaire.

Model	Personalization	F1- macro	Std	t	p	Signif.
Baseline	-	0.419	0.021	-	-	-
KNN	Broad context	0.634	0.067	8.597	0.000	***
	General	0.544	0.058	6.040	0.000	***
	Personality	0.583	0.067	6.660	0.000	***
MLP	Broad context	0.624	0.065	8.699	0.000	***
	General	0.580	0.070	6.586	0.000	***
	Personality	0.596	0.067	7.323	0.000	***
Random Forest	Broad context	0.513	0.067	4.446	0.000	***
	General	0.498	0.067	3.898	0.001	***
	Personality	0.516	0.072	4.170	0.000	***
SVM	Broad context	0.578	0.069	6.831	0.000	***
	General	0.428	0.036	0.987	0.326	ns
	Personality	0.473	0.057	3.370	0.002	**

Table C.7: Results for classification of next day morning stress questionnaire.

Model	Personalization	F1- macro	Std	t	p	Signif.
Baseline	-	0.419	0.018	-	-	-
KNN	Broad context	0.772	0.054	16.899	0.000	***
	General	0.764	0.068	14.225	0.000	***
	Personality	0.755	0.063	14.699	0.000	***
MLP	Broad context	0.766	0.057	16.380	0.000	***
	General	0.726	0.062	14.107	0.000	***
	Personality	0.736	0.062	13.798	0.000	***
Random Forest	Broad context	0.735	0.062	14.123	0.000	***
	General	0.732	0.064	13.664	0.000	***
	Personality	0.740	0.067	13.600	0.000	***
SVM	Broad context	0.757	0.057	15.815	0.000	***
	General	0.754	0.066	14.124	0.000	***
	Personality	0.748	0.067	13.456	0.000	***

Table C.8: Results for classification of next day morning composure questionnaire.

Model	Personalization	F1- macro	Std	t	p	Signif.
Baseline	-	0.390	0.023	-	-	-
KNN	Broad context	0.742	0.057	16.085	0.000	***
	General	0.704	0.058	14.166	0.000	***
	Personality	0.729	0.058	15.567	0.000	***
MLP	Broad context	0.739	0.059	14.957	0.000	***
	General	0.661	0.056	12.741	0.000	***
	Personality	0.698	0.058	14.552	0.000	***
Random Forest	Broad context	0.670	0.063	12.408	0.000	***
	General	0.674	0.056	13.123	0.000	***
	Personality	0.691	0.062	13.276	0.000	***
SVM	Broad context	0.722	0.058	15.819	0.000	***
	General	0.713	0.060	14.897	0.000	***
	Personality	0.714	0.060	14.909	0.000	***

Table C.9: Results for regression of same day evening health questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.231	0.013	-	-	-
KNN	Broad context	0.197	0.015	-7.968	0.000	***
	General	0.220	0.012	-3.372	0.003	**
	Personality	0.207	0.014	-7.574	0.000	***
MLP	Broad context	0.221	0.016	-1.975	0.102	ns
	General	0.256	0.017	4.295	0.000	***
	Personality	0.236	0.015	0.987	0.326	ns
Random Forest	Broad context	0.197	0.011	-12.329	0.000	***
	General	0.208	0.010	-9.147	0.000	***
	Personality	0.203	0.012	-10.783	0.000	***
SVM	Broad context	0.188	0.013	-11.894	0.000	***
	General	0.208	0.012	-8.045	0.000	***
	Personality	0.203	0.014	-9.216	0.000	***

Table C.10: Results for regression of same day evening mood questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.208	0.009	-	-	-
KNN	Broad context	0.189	0.012	-5.083	0.000	***
	General	0.193	0.011	-5.073	0.000	***
	Personality	0.193	0.011	-4.807	0.000	***
MLP	Broad context	0.209	0.015	0.187	0.852	ns
	General	0.213	0.017	0.842	0.804	ns
	Personality	0.222	0.014	2.748	0.021	*
Random Forest	Broad context	0.176	0.010	-13.047	0.000	***
	General	0.179	0.009	-12.175	0.000	***
	Personality	0.182	0.010	-11.645	0.000	***
SVM	Broad context	0.173	0.010	-10.290	0.000	***
	General	0.177	0.010	-10.249	0.000	***
	Personality	0.180	0.010	-9.093	0.000	***

Table C.11: Results for regression of same day evening overwhelm questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.304	0.012	-	-	-
KNN	Broad context	0.246	0.014	-11.893	0.000	***
	General	0.267	0.014	-8.189	0.000	***
	Personality	0.255	0.014	-10.848	0.000	***
MLP	Broad context	0.280	0.017	-3.658	0.001	**
	General	0.301	0.019	-0.365	0.716	ns
	Personality	0.293	0.019	-1.652	0.203	ns
Random Forest	Broad context	0.240	0.013	-16.182	0.000	***
Forest	General	0.247	0.013	-14.662	0.000	***
	Personality	0.244	0.013	-15.182	0.000	***
SVM	Broad context	0.234	0.013	-15.465	0.000	***
	General	0.248	0.015	-12.269	0.000	***
	Personality	0.243	0.014	-14.067	0.000	***

Table C.12: Results for regression of same day evening unpredictability questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.290	0.012	-	-	-
KNN	Broad context	0.229	0.015	-12.601	0.000	***
	General	0.262	0.013	-6.031	0.000	***
	Personality	0.240	0.013	-10.970	0.000	***
MLP	Broad context	0.263	0.019	-3.929	0.000	***
	General	0.302	0.019	1.734	0.172	ns
	Personality	0.282	0.018	-1.178	0.242	ns
Random Forest	Broad context	0.234	0.011	-16.707	0.000	***
	General	0.251	0.012	-10.957	0.000	***
	Personality	0.241	0.011	-14.127	0.000	***
SVM	Broad context	0.221	0.014	-15.262	0.000	***
	General	0.249	0.014	-9.374	0.000	***
	Personality	0.241	0.013	-12.712	0.000	***

Table C.13: Results for classification of same day evening health questionnaire.

Model	Personalization	F1-macro	Std	t	p	Signif.
Baseline	-	0.465	0.010	-	-	-
KNN	Broad context	0.665	0.073	7.839	0.000	***
	General	0.556	0.065	4.003	0.000	***
	Personality	0.652	0.079	6.994	0.000	***
MLP	Broad context	0.670	0.067	8.755	0.000	***
	General	0.614	0.067	6.384	0.000	***
	Personality	0.653	0.076	7.000	0.000	***
Random Forest	Broad context	0.561	0.074	3.922	0.000	***
Forest	General	0.514	0.058	2.623	0.010	*
	Personality	0.519	0.062	2.667	0.018	*
SVM	Broad context	0.585	0.065	5.412	0.000	***
	General	0.465	0.010	0.000	0.000	***
	Personality	0.573	0.064	5.031	0.000	***

Table C.14: Results for classification of same day evening mood questionnaire.

Model	Personalization	F1- macro	Std	t	p	Signif.
Baseline	-	0.443	0.011	-	-	-
KNN	Broad context	0.685	0.060	11.615	0.000	***
	General	0.624	0.051	9.792	0.000	***
	Personality	0.671	0.056	11.634	0.000	***
MLP	Broad context	0.706	0.058	13.024	0.000	***
	General	0.685	0.056	12.049	0.000	***
	Personality	0.697	0.051	13.902	0.000	***
Random Forest	Broad context	0.689	0.060	12.033	0.000	***
	General	0.673	0.067	9.962	0.000	***
	Personality	0.611	0.070	7.028	0.000	***
SVM	Broad context	0.622	0.064	8.219	0.000	***
	General	0.662	0.066	9.656	0.000	***
	Personality	0.644	0.060	9.947	0.000	***

Table C.15: Results for classification of same day evening overwhelm questionnaire.

Model	Personalization	F1- macro	Std	t	p	Signif.
Baseline	-	0.436	0.011	-	-	-
KNN	Broad context	0.683	0.055	12.721	0.000	***
	General	0.684	0.048	14.023	0.000	***
	Personality	0.685	0.053	13.344	0.000	***
MLP	Broad context	0.701	0.053	13.537	0.000	***
	General	0.674	0.045	14.354	0.000	***
	Personality	0.694	0.055	13.059	0.000	***
Random Forest	Broad context	0.666	0.057	11.912	0.000	***
	General	0.650	0.056	11.586	0.000	***
	Personality	0.653	0.060	10.787	0.000	***
SVM	Broad context	0.662	0.053	12.244	0.000	***
	General	0.704	0.053	14.632	0.000	***
	Personality	0.701	0.049	15.483	0.000	***

Table C.16: Results for classification of same day evening unpredictability questionnaire.

Model	Personalization	F1-macro	Std	t	p	Signif.
Baseline	-	0.436	0.010	-	-	-
KNN	Broad context	0.713	0.049	16.051	0.000	***
	General	0.661	0.055	11.762	0.000	***
	Personality	0.670	0.055	12.067	0.000	***
MLP	Broad context	0.712	0.052	14.823	0.000	***
	General	0.623	0.059	9.053	0.000	***
	Personality	0.684	0.048	13.786	0.000	***
Random Forest	Broad context	0.613	0.057	9.210	0.000	***
	General	0.574	0.058	7.045	0.000	***
	Personality	0.577	0.056	7.441	0.000	***
SVM	Broad context	0.634	0.063	9.211	0.000	***
	General	0.572	0.060	6.931	0.000	***
	Personality	0.625	0.062	8.988	0.000	***

Table C.17: Results for regression of daily emotions morning valence questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.497	0.004	-	-	-
KNN	Broad context	0.518	0.029	2.152	0.135	ns
	General	0.530	0.028	3.403	0.006	**
	Personality	0.535	0.022	4.869	0.000	***
MLP	Broad context	0.608	0.043	7.523	0.000	***
	General	0.619	0.039	8.973	0.000	***
	Personality	0.621	0.045	7.962	0.000	***
Random Forest	Broad context	0.496	0.015	-0.231	1.636	ns
	General	0.500	0.014	0.674	1.505	ns
	Personality	0.497	0.014	0.128	0.898	ns
SVM	Broad context	0.535	0.033	3.381	0.005	**
	General	0.537	0.031	3.753	0.002	**
	Personality	0.536	0.030	3.739	0.002	**

Table C.18: Results for regression of daily emotions morning arousal questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.491	0.009	-	-	-
KNN	Broad context	0.489	0.032	-0.233	2.450	ns
	General	0.525	0.026	4.251	0.000	***
	Personality	0.508	0.025	1.814	0.436	ns
MLP	Broad context	0.570	0.040	5.665	0.000	***
	General	0.611	0.049	7.431	0.000	***
	Personality	0.601	0.039	8.204	0.000	***
Random Forest	Broad context	0.470	0.021	-3.143	0.018	*
	General	0.485	0.018	-1.196	1.174	ns
	Personality	0.478	0.019	-2.151	0.237	ns
SVM	Broad context	0.490	0.035	-0.115	0.909	ns
	General	0.493	0.031	0.153	1.757	ns
	Personality	0.494	0.029	0.317	3.007	ns

Table C.19: Results for regression of daily emotions afternoon valence questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.472	0.016	-	-	-
KNN	Broad context	0.489	0.028	1.932	0.281	ns
	General	0.503	0.028	3.822	0.002	**
	Personality	0.492	0.028	2.566	0.083	ns
MLP	Broad context	0.577	0.038	7.960	0.000	***
	General	0.591	0.042	8.437	0.000	***
	Personality	0.573	0.044	6.971	0.000	***
Random Forest	Broad context	0.461	0.019	-2.022	0.275	ns
Forest	General	0.472	0.017	-0.040	0.968	ns
	Personality	0.466	0.017	-1.220	0.450	ns
SVM	Broad context	0.483	0.033	1.237	0.657	ns
	General	0.494	0.034	2.940	0.033	*
	Personality	0.485	0.033	1.613	0.439	ns

Table C.20: Results for regression of daily emotions afternoon arousal questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.441	0.022	-	-	-
KNN	Broad context	0.448	0.034	0.578	2.259	ns
	General	0.454	0.027	1.871	0.450	ns
	Personality	0.445	0.032	0.405	2.058	ns
MLP	Broad context	0.527	0.044	6.019	0.000	***
	General	0.538	0.038	7.464	0.000	***
	Personality	0.528	0.042	6.278	0.000	***
Random Forest	Broad context	0.436	0.027	-0.852	2.379	ns
	General	0.443	0.027	0.340	1.470	ns
	Personality	0.440	0.027	-0.107	0.915	ns
SVM	Broad context	0.449	0.036	0.811	2.096	ns
	General	0.463	0.038	3.505	0.006	**
	Personality	0.458	0.038	2.734	0.059	ns

Table C.21: Results for regression of daily emotions evening valence questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.383	0.033	-	-	-
KNN	Broad context	0.393	0.038	1.230	0.665	ns
	General	0.412	0.036	4.048	0.001	***
	Personality	0.410	0.038	3.331	0.009	**
MLP	Broad context	0.449	0.043	5.790	0.000	***
	General	0.459	0.042	6.433	0.000	***
	Personality	0.454	0.046	6.143	0.000	***
Random Forest	Broad context	0.381	0.032	-0.326	1.490	ns
Forest	General	0.399	0.033	3.625	0.004	**
	Personality	0.392	0.033	1.986	0.199	ns
SVM	Broad context	0.381	0.040	-0.218	0.828	ns
	General	0.392	0.041	3.175	0.012	*
	Personality	0.390	0.042	2.244	0.135	ns

Table C.22: Results for regression of daily emotions evening arousal questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.393	0.033	-	-	-
KNN	Broad context	0.402	0.039	1.081	0.847	ns
	General	0.433	0.037	5.670	0.000	***
	Personality	0.426	0.039	4.407	0.000	***
MLP	Broad context	0.448	0.042	5.280	0.000	***
	General	0.498	0.045	8.310	0.000	***
	Personality	0.471	0.047	5.887	0.000	***
Random Forest	Broad context	0.393	0.030	-0.038	0.969	ns
	General	0.409	0.031	3.147	0.013	*
	Personality	0.403	0.031	1.740	0.340	ns
SVM	Broad context	0.395	0.041	0.393	1.391	ns
	General	0.405	0.041	3.758	0.002	**
	Personality	0.404	0.042	3.016	0.016	*

Table C.23: Results for classification of daily emotions morning valence questionnaire.

Model	Personalization	F1-macro	Std	t	p	Signif.
Baseline	-	0.497	0.004	-	-	-
KNN	Broad context	0.565	0.051	3.787	0.003	**
	General	0.550	0.055	2.787	0.038	*
	Personality	0.541	0.052	2.437	0.050	*
MLP	Broad context	0.551	0.054	2.854	0.037	*
	General	0.531	0.057	1.726	0.088	ns
	Personality	0.535	0.052	2.106	0.075	ns
Random Forest	Broad context	0.557	0.056	3.041	0.033	*
Forest	General	0.552	0.061	2.568	0.047	*
	Personality	0.557	0.062	2.775	0.033	*
SVM	Broad context	0.547	0.050	2.878	0.039	*
	General	0.559	0.061	2.915	0.040	*
	Personality	0.557	0.057	3.016	0.033	*

Table C.24: Results for classification of daily emotions morning arousal questionnaire.

Model	Personalization	F1-macro	Std	t	p	Signif.
Baseline	-	0.491	0.009	-	-	-
KNN	Broad context	0.618	0.055	6.487	0.000	***
	General	0.518	0.055	1.340	0.183	ns
	Personality	0.584	0.056	4.781	0.000	***
MLP	Broad context	0.604	0.052	6.026	0.000	***
	General	0.539	0.054	2.519	0.027	*
	Personality	0.559	0.053	3.543	0.002	**
Random Forest	Broad context	0.634	0.056	7.110	0.000	***
SVM	General	0.576	0.053	4.413	0.000	***
	Personality	0.590	0.056	4.953	0.000	***
SVM	Broad context	0.645	0.054	7.896	0.000	***
	General	0.597	0.051	5.680	0.000	***
	Personality	0.608	0.054	6.004	0.000	***

Table C.25: Results for classification of daily emotions afternoon valence questionnaire.

Model	Personalization	F1-macro	Std	t	p	Signif.
Baseline	-	0.472	0.016	-	-	-
KNN	Broad context	0.579	0.049	6.021	0.000	***
	General	0.542	0.055	3.417	0.007	**
	Personality	0.559	0.059	4.004	0.001	**
MLP	Broad context	0.563	0.057	4.286	0.000	***
	General	0.538	0.061	3.034	0.022	*
	Personality	0.556	0.057	4.218	0.001	***
Random Forest	Broad context	0.543	0.062	2.991	0.018	*
SVM	General	0.491	0.054	0.898	0.743	ns
	Personality	0.513	0.053	1.932	0.225	ns
SVM	Broad context	0.543	0.060	3.021	0.019	*
	General	0.437	0.051	-1.579	0.353	ns
	Personality	0.468	0.050	-0.167	0.868	ns

Table C.26: Results for classification of daily emotions afternoon arousal questionnaire.

Model	Personalization	F1-macro	Std	t	p	Signif.
Baseline	-	0.441	0.022	-	-	-
KNN	Broad context	0.617	0.058	8.482	0.000	***
	General	0.558	0.058	5.086	0.000	***
	Personality	0.590	0.054	7.243	0.000	***
MLP	Broad context	0.592	0.063	6.456	0.000	***
	General	0.546	0.056	5.046	0.000	***
	Personality	0.570	0.054	6.091	0.000	***
Random Forest	Broad context	0.496	0.053	2.516	0.081	ns
SVM	General	0.459	0.049	0.823	0.825	ns
	Personality	0.466	0.048	1.168	0.737	ns
SVM	Broad context	0.432	0.033	-0.544	0.587	ns
	General	0.421	0.016	-1.495	0.691	ns
	Personality	0.421	0.016	-1.495	0.553	ns

Table C.27: Results for classification of daily emotions evening valence questionnaire.

Model	Personalization	F1-macro	Std	t	p	Signif.
Baseline	-	0.383	0.033	-	-	-
KNN	Broad context	0.585	0.068	7.429	0.000	***
	General	0.497	0.058	4.523	0.000	***
	Personality	0.541	0.065	5.781	0.000	***
MLP	Broad context	0.578	0.060	8.307	0.000	***
	General	0.529	0.060	5.858	0.000	***
	Personality	0.554	0.065	6.192	0.000	***
Random Forest	Broad context	0.456	0.027	4.068	0.000	***
SVM	General	0.449	0.018	4.066	0.000	***
	Personality	0.460	0.032	3.897	0.000	***
SVM	Broad context	0.449	0.012	4.229	0.000	***
	General	0.450	0.012	4.234	0.000	***
	Personality	0.450	0.012	4.234	0.000	***

Table C.28: Results for classification of daily emotions evening arousal questionnaire.

Model	Personalization	F1-macro	Std	t	p	Signif.
Baseline	-	0.393	0.033	-	-	-
KNN	Broad context	0.576	0.066	6.419	0.000	***
	General	0.477	0.044	3.872	0.001	**
	Personality	0.502	0.054	4.475	0.000	***
MLP	Broad context	0.564	0.060	7.465	0.000	***
	General	0.508	0.056	4.763	0.000	***
	Personality	0.528	0.054	5.849	0.000	***
Random Forest	Broad context	0.452	0.025	3.437	0.004	**
	General	0.452	0.027	3.260	0.002	**
	Personality	0.454	0.026	3.531	0.004	**
SVM	Broad context	0.446	0.013	3.339	0.005	**
	General	0.446	0.013	3.339	0.004	**
	Personality	0.446	0.013	3.339	0.002	**

Table C.29: Results for regression of momentary emotions valence questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.226	0.006	-	-	-
KNN	Broad context	0.232	0.008	2.830	0.017	*
	General	0.244	0.007	9.258	0.000	***
	Personality	0.234	0.008	3.711	0.002	**
MLP	Broad context	0.254	0.022	3.589	0.002	**
	General	0.222	0.006	-3.731	0.002	**
	Personality	0.262	0.016	7.093	0.000	***
Random Forest	Broad context	0.215	0.007	-8.363	0.000	***
	General	0.223	0.006	-2.264	0.026	*
	Personality	0.216	0.007	-7.865	0.000	***
SVM	Broad context	0.216	0.008	-4.458	0.000	***
	General	0.222	0.007	-2.509	0.027	*
	Personality	0.217	0.007	-5.013	0.000	***

Table C.30: Results for regression of momentary emotions arousal questionnaire.

Model	Personalization	RMSE	Std	t	p	Signif.
Baseline	-	0.196	0.007	-	-	-
KNN	Broad context	0.197	0.008	0.301	0.764	ns
	General	0.211	0.008	9.797	0.000	***
	Personality	0.199	0.008	1.229	0.444	ns
MLP	Broad context	0.209	0.020	1.920	0.173	ns
	General	0.193	0.007	-2.565	0.047	*
	Personality	0.215	0.016	3.795	0.001	**
Random Forest	Broad context	0.181	0.007	-11.006	0.000	***
	General	0.193	0.006	-4.645	0.000	***
	Personality	0.183	0.006	-12.590	0.000	***
SVM	Broad context	0.182	0.007	-8.364	0.000	***
	General	0.192	0.007	-5.240	0.000	***
	Personality	0.186	0.007	-7.482	0.000	***

Table C.31: Results for classification of momentary emotions intense emotions questionnaire.

Model	Personalization	F1-macro	Std	t	p	Signif.
Baseline	-	0.461	0.013	-	-	-
KNN	Broad context	0.699	0.028	23.641	0.000	***
	General	0.494	0.032	2.586	0.022	*
	Personality	0.671	0.030	20.200	0.000	***
MLP	Broad context	0.680	0.029	19.200	0.000	***
	General	0.514	0.032	4.165	0.000	***
	Personality	0.663	0.033	16.665	0.000	***
Random Forest	Broad context	0.695	0.028	22.090	0.000	***
Forest	General	0.451	0.020	-1.129	0.262	ns
	Personality	0.508	0.029	3.713	0.001	**
SVM	Broad context	0.708	0.029	22.932	0.000	***
	General	0.408	0.011	-6.416	0.000	***
	Personality	0.413	0.018	-4.777	0.000	***

Table C.32: Results for classification of momentary emotions valence questionnaire.

Model	Personalization	F1- macro	Std	t	p	Signif.
Baseline	-	0.468	0.009	-	-	-
KNN	Broad context	0.601	0.029	12.010	0.000	***
	General	0.518	0.027	4.907	0.000	***
	Personality	0.590	0.029	11.428	0.000	***
MLP	Broad context	0.580	0.026	11.536	0.000	***
	General	0.526	0.028	5.555	0.000	***
	Personality	0.582	0.027	11.395	0.000	***
Random Forest	Broad context	0.531	0.028	5.678	0.000	***
	General	0.469	0.022	0.110	0.913	ns
	Personality	0.492	0.026	2.273	0.101	ns
SVM	Broad context	0.573	0.032	8.472	0.000	***
	General	0.452	0.026	-1.425	0.471	ns
	Personality	0.482	0.030	1.222	0.449	ns

Table C.33: Results for classification of momentary emotions arousal questionnaire.

Model	Personalization	F1- macro	Std	t	p	Signif.
Baseline	-	0.398	0.019	-	-	-
KNN	Broad context	0.570	0.034	12.570	0.000	***
	General	0.515	0.026	9.924	0.000	***
	Personality	0.571	0.035	12.227	0.000	***
MLP	Broad context	0.589	0.032	15.929	0.000	***
	General	0.532	0.031	10.605	0.000	***
	Personality	0.580	0.033	14.771	0.000	***
Random Forest	Broad context	0.481	0.022	7.243	0.000	***
	General	0.461	0.016	6.086	0.000	***
	Personality	0.468	0.019	6.338	0.000	***
SVM	Broad context	0.484	0.033	6.150	0.000	***
	General	0.445	0.008	5.065	0.000	***
	Personality	0.445	0.008	5.065	0.000	***

C.2.2 Pairwise comparisons of modeling strategies

Table C.34: Pairwise comparisons of modeling strategies for regression of next day morning questionnaires.

Task	Approach 1	Approach 2	t	p	Signif.
Composure	General	Broad context	3.127	0.007	**
		Personality	2.872	0.010	**
Rest	Personality	Broad context	1.328	0.187	ns
	General	Broad context	1.865	0.195	ns
Sleep Quality	Personality	Broad context	1.620	0.217	ns
		General	Broad context	1.899	0.182
	Personality	Personality	1.292	0.399	ns
Stress	Personality	Broad context	1.161	0.249	ns
	General	Broad context	-0.838	0.404	ns
	Personality	Personality	1.326	0.376	ns
	Personality	Broad context	-1.862	0.197	ns

Table C.35: Pairwise comparisons of modeling strategies for classification of next day morning questionnaires.

Task	Approach 1	Approach 2	t	p	Signif.
Composure	General	Broad context	-2.175	0.064	ns
		Personality	-2.288	0.073	ns
Rest	Personality	Broad context	-0.801	0.425	ns
	General	Broad context	-4.557	0.000	***
Sleep Quality	Personality	Broad context	-2.349	0.021	*
		General	Broad context	-3.640	0.001
	Personality	Personality	-2.441	0.049	*
Stress	Personality	Broad context	-0.715	0.476	ns
	General	Broad context	-2.332	0.044	*
	Personality	Personality	-0.915	0.724	ns
	Personality	Broad context	-0.082	0.935	ns
	Personality	Broad context	-0.981	0.987	ns

Table C.36: Pairwise comparisons of modeling strategies for regression of same day evening questionnaires.

Task	Approach 1	Approach 2	t	p	Signif.
Health	General	Broad context	7.304	0.000	***
		Personality	4.807	0.000	***
Mood	Personality	Broad context	5.063	0.000	***
	General	Broad context	1.323	0.189	ns
Overwhelm	General	Personality	-1.767	0.161	ns
		Broad context	4.205	0.000	***
	Personality	Broad context	5.535	0.000	***
Unpredictability	General	Personality	3.327	0.001	**
		Broad context	4.081	0.000	***
	Personality	Broad context	7.894	0.000	***
	General	Personality	5.260	0.000	***
		Personality	Broad context	5.707	0.000

Table C.37: Pairwise comparisons of modeling strategies for classification of same day evening questionnaires.

Task	Approach 1	Approach 2	t	p	Signif.
Health	General	Broad context	-5.132	0.000	***
		Personality	-4.047	0.000	***
Mood	Personality	Broad context	-1.851	0.067	ns
	General	Broad context	-1.018	0.622	ns
Overwhelm	General	Personality	0.421	0.675	ns
		Broad context	-1.705	0.274	ns
	Personality	Broad context	-0.012	0.991	ns
Unpredictability	General	Personality	-0.551	1.166	ns
		Broad context	0.575	1.700	ns
	Personality	Broad context	-4.312	0.000	***
	General	Personality	-2.879	0.010	**
		Personality	Broad context	-2.801	0.006

Table C.38: Pairwise comparisons of modeling strategies for regression of daily emotions questionnaires.

Task	Approach 1	Approach 2	t	p	Signif.
Afternoon Arousal	General	Broad context	1.459	0.443	ns
	Personality	Broad context	0.507	0.614	ns
Afternoon Valence	General	Broad context	1.562	0.243	ns
	Personality	Broad context	0.186	0.853	ns
Evening Arousal	General	Broad context	4.927	0.000	***
	Personality	Broad context	2.145	0.034	*
	Personality	Broad context	3.484	0.001	**
Evening Valence	General	Broad context	2.100	0.115	ns
	Personality	Broad context	0.777	0.439	ns
	Personality	Broad context	1.936	0.111	ns
Morning Arousal	General	Broad context	2.754	0.021	*
	Personality	Broad context	1.240	0.218	ns
	Personality	Broad context	2.319	0.045	*
Morning Valence	General	Broad context	0.845	0.800	ns
	Personality	Broad context	-0.096	0.924	ns
	Personality	Broad context	1.245	0.648	ns

Table C.39: Pairwise comparisons of modeling strategies for classification of daily emotions questionnaires.

Task	Approach 1	Approach 2	t	p	Signif.
Afternoon Arousal	General	Broad context	-3.144	0.007	**
	Personality	Broad context	-2.360	0.040	*
Afternoon Valence	General	Broad context	-3.505	0.002	**
	Personality	Broad context	-2.638	0.019	*
Evening Arousal	General	Broad context	-3.576	0.002	**
	Personality	Broad context	-2.963	0.008	**
Evening Valence	General	Broad context	-3.319	0.004	**
	Personality	Broad context	-2.053	0.085	ns
Morning Arousal	General	Broad context	-5.155	0.000	***
	Personality	Broad context	-2.645	0.010	**
Morning Valence	General	Broad context	-0.548	1.169	ns
	Personality	Broad context	-0.688	1.480	ns

Table C.40: Pairwise comparisons of modeling strategies for regression of momentary emotions questionnaires.

Task	Approach 1	Approach 2	t	p	Signif.
Arousal	General	Broad context	2.318	0.067	ns
	Personality	Broad context	1.667	0.197	ns
Valence	General	Broad context	-0.608	0.545	ns
	Personality	Broad context	1.244	0.433	ns

Table C.41: Pairwise comparisons of modeling strategies for classification of momentary emotions questionnaires.

Task	Approach 1	Approach 2	t	p	Signif.
Arousal	General	Broad context	-5.601	0.000	***
		Personality	-4.751	0.000	***
Intense Emotions	Personality	Broad context	-2.853	0.005	**
	General	Broad context	-23.935	0.000	***
		Personality	-13.222	0.000	***
Valence	Personality	Broad context	-22.003	0.000	***
	General	Broad context	-10.187	0.000	***
		Personality	-7.074	0.000	***
	Personality	Broad context	-7.103	0.000	***

Appendix D

EPIC COMPETITION SUBMISSIONS

Table D.1: Details of submissions to Emotion Physiology and Experience Collaboration (EPiC) challenge (from [32]). ? - deduced by us.

Team id	Approach	Comment
Team 1	<ul style="list-style-type: none"> • Combining multiple physiological signals into one, using their variance as weights • Person-specific models in across-time, across-emotion and across-induction scenarios? • MLP models used in all scenarios • Predicting arousal and valence levels at the same time 	Non-rigorous approach – mixing data between folds
Team 2	<ul style="list-style-type: none"> • Models utilizing convolutional and recurrent layers (TCN-LSTM) [247] • Predicting arousal and valence levels separately 	Intended (rigorous) approach to validation
Team 3	<ul style="list-style-type: none"> • Models trained in the smallest possible context <ul style="list-style-type: none"> – Separate models for each person-video pair in across-time validation scenario – Video-specific models in across-subject scenario – Models trained on selected most informative videos in across-emotion scenario – Separate models for each video in across-induction scenario • Autogluon (AutoML framework) used to train models and choose the best one: <ul style="list-style-type: none"> – MLP – Tree-based algorithms (Random Forest, CatBoost, GBM, GBMXT, GBMLarge) – KNN • Predicting arousal and valence levels separately 	Intended (rigorous) approach to validation

Team id	Approach	Comment
Team 4	<ul style="list-style-type: none"> • Tree-based algorithms <ul style="list-style-type: none"> – Separate models for each person-video pair in across-time validation scenario – Video-specific models in across-subject scenario – Subject-specific models in across-emotion scenario and across-induction scenario • Tree based algorithms (Random Forest and XGBoost) • Predicting arousal and valence levels separately 	Intended (rigorous) approach to validation
Team 5	<ul style="list-style-type: none"> • General models predicting arousal and valence levels at the same time? • Multivariate Time Series Transformers in across-time and across-subject validation scenarios • XGBoost models in across-emotion scenario and across-induction scenarios 	Non-rigorous approach – mixing data between folds
Team 6	<ul style="list-style-type: none"> • General models predicting arousal and valence levels separately • In each validation scenario, three algorithms were tested (ElasticNet, Random Forest, SVM) and the best one was selected 	Intended (rigorous) approach to validation
Team 7	<ul style="list-style-type: none"> • General models simultaneously predicting arousal and valence levels • FEDformer (Frequency Enhanced Decomposed Transformer) algorithm used, with additional model for prediction smoothing 	Non-rigorous approach – mixing data between folds?
Team 8	<ul style="list-style-type: none"> • General models simultaneously predicting arousal and valence levels • Models consisting of LSTM layers • In across-time validation scenario, models were provided with information about predicted subject and video 	Intended (rigorous) approach to validation

Team id	Approach	Comment
Team 9	<ul style="list-style-type: none"> • Only ECG signal used • Models trained in the smallest possible context <ul style="list-style-type: none"> – Separate models for each person-video pair in across-time validation scenario – Video-specific models in across-subject scenario – Subject-specific models in across-emotion scenario and across-induction scenario • Predicting arousal and valence levels separately • Deep transformer-based network used 	
Team 10	<ul style="list-style-type: none"> • General models predicting arousal and valence levels separately • Deep pre-trained model used <ul style="list-style-type: none"> – S4 state-space architecture – Layers for ECG signal processing was pre-trained on TILES [231] dataset – Layers for other signals were trained in supervised manner on competition dataset 	Intended (rigorous) approach to validation
Team 11	<ul style="list-style-type: none"> • General models for predicting arousal and valence levels simultaneously • Only decision trees used 	Non-rigorous approach – mixing data between folds
Team 12	<ul style="list-style-type: none"> • General models predicting arousal and valence levels separately • Deep transformer-based network used 	Intended (rigorous) approach to validation?

Table D.2: Competition results in across-time validation scenario (based on [32]).

Team	MAE	diff (mean)	Z (mean)	p (mean)	diff (random)	Z (random)	p (random)
Baseline (mean)	1.539	-	-	-	-	-	-
Baseline (random)	2.700	-	-	-	-	-	-
Team 1	1.657	-0.118	-34.668	0.000	-0.118	306.242	0.000
Team 2	1.584	-0.046	-12.911	0.000	-0.046	316.421	0.000
Team 3	0.876	0.663	196.675	0.000	0.663	540.834	0.000
Team 4	0.807	0.732	214.015	0.000	0.732	553.365	0.000
Team 5	1.033	0.506	154.009	0.000	0.506	507.092	0.000
Team 6	1.646	-0.107	-31.249	0.000	-0.107	307.931	0.000
Team 7	1.685	-0.146	-41.942	0.000	-0.146	291.599	0.000
Team 8	1.311	0.228	63.337	0.000	0.228	385.288	0.000
Team 9	0.937	0.602	177.776	0.000	0.602	520.448	0.000
Team 10	1.349	0.190	53.221	0.000	0.190	377.956	0.000
Team 11	1.540	-0.001	-0.171	0.984	-0.001	320.925	0.000
Team 12	1.521	0.018	5.367	0.000	0.018	344.229	0.000

Table D.3: Competition results in across-subject validation scenario (based on [32]).

Team	MAE	diff (mean)	Z (mean)	p (mean)	diff (random)	Z (random)	p (random)
Baseline (mean)	1.249	-	-	-	-	-	-
Baseline (random)	2.568	-	-	-	-	-	-
Team 1	1.240	0.009	2.541	0.030	0.009	389.845	0.000
Team 2	1.393	-0.144	-40.972	0.000	-0.144	333.179	0.000
Team 3	1.026	0.223	66.207	0.000	0.223	457.202	0.000
Team 4	1.218	0.031	9.024	0.000	0.031	394.556	0.000
Team 5	0.759	0.490	149.071	0.000	0.490	550.204	0.000
Team 6	1.029	0.220	64.279	0.000	0.220	449.617	0.000
Team 7	1.505	-0.256	-73.630	0.000	-0.256	305.303	0.000
Team 8	2.085	-0.837	-232.000	0.000	-0.837	133.765	0.000
Team 9	0.965	0.283	83.644	0.000	0.283	472.950	0.000
Team 10	1.332	-0.084	-23.397	0.000	-0.084	345.531	0.000
Team 11	1.072	0.177	48.987	0.000	0.177	413.780	0.000
Team 12	1.276	-0.027	-7.912	0.000	-0.027	377.066	0.000

Table D.4: Competition results in across-emotion validation scenario (based on [32]).

Team	MAE	diff (mean)	Z (mean)	p (mean)	diff (random)	Z (random)	p (random)
Baseline (mean)	1.404	-	-	-	-	-	-
Baseline (random)	2.561	-	-	-	-	-	-
Team 1	0.972	0.432	126.742	0.000	0.432	466.711	0.000
Team 2	1.583	-0.179	-50.735	0.000	-0.179	277.689	0.000
Team 3	1.359	0.045	13.254	0.000	0.045	356.464	0.000
Team 4	1.513	-0.109	-31.981	0.000	-0.109	306.433	0.000
Team 5	0.700	0.704	214.122	0.000	0.704	566.230	0.000
Team 6	1.383	0.021	6.014	0.000	0.021	344.258	0.000
Team 7	1.536	-0.132	-38.014	0.000	-0.132	294.607	0.000
Team 8	1.660	-0.256	-70.982	0.000	-0.256	250.081	0.000
Team 9	1.521	-0.117	-34.530	0.000	-0.117	307.196	0.000
Team 10	1.436	-0.033	-9.156	0.000	-0.033	314.683	0.000
Team 11	1.060	0.344	95.156	0.000	0.344	415.366	0.000
Team 12	1.444	-0.040	-11.656	0.000	-0.040	326.272	0.000

Table D.5: Competition results in across-induction validation scenario (based on [32]).

Team	MAE	diff (mean)	Z (mean)	p (mean)	diff (random)	Z (random)	p (random)
Baseline (mean)	1.253	-	-	-	-	-	-
Baseline (random)	2.572	-	-	-	-	-	-
Team 1	0.942	0.310	91.122	0.000	0.310	478.457	0.000
Team 2	1.554	-0.301	-85.483	0.000	-0.301	288.698	0.000
Team 3	1.255	-0.003	-0.754	0.731	-0.003	390.272	0.000
Team 4	1.397	-0.144	-42.146	0.000	-0.144	343.416	0.000
Team 5	0.742	0.510	155.253	0.000	0.510	556.418	0.000
Team 6	1.495	-0.242	-70.789	0.000	-0.242	314.580	0.000
Team 7	1.305	-0.053	-15.137	0.000	-0.053	363.826	0.000
Team 8	1.460	-0.208	-57.639	0.000	-0.208	308.154	0.000
Team 9	1.336	-0.083	-24.604	0.000	-0.083	364.732	0.000
Team 10	2.054	-0.801	-224.113	0.000	-0.801	144.844	0.000
Team 11	1.021	0.232	64.203	0.000	0.232	429.025	0.000
Team 12	1.297	-0.044	-12.807	0.000	-0.044	372.202	0.000

Table D.6: Results for the three teams selected to partake in additional testing (based on [32]).

Scenario	Team	Team 3	Team 4	Team 9
Across-time	MAE	0.876	0.807	0.937
	MAE (random)	1.024	0.956	0.986
	diff	-0.148	-0.150	-0.049
	Z	-58.318	-56.907	-18.701
	p	0.000	0.000	0.000
Across-subject	MAE	1.026	1.218	0.965
	MAE (random)	1.092	1.285	0.984
	diff	-0.067	-0.067	-0.018
	Z	-26.305	-25.409	-6.969
	p	0.000	0.000	0.000
Across-emotion	MAE	1.359	1.513	1.521
	MAE (random)	1.687	1.516	1.650
	diff	-0.328	-0.003	-0.129
	Z	-129.180	-1.066	-49.300
	p	0.000	0.287	0.000
Across-induction	MAE	1.255	1.397	1.336
	MAE (random)	1.275	1.493	1.478
	diff	-0.020	-0.096	-0.142
	Z	-7.992	-36.607	-54.263
	p	0.000	0.000	0.000

Appendix E

PERSONALIZED EMOTION RECOGNITION

E.1 Proof for z-score equality

Population data $x = \{x_1, x_2, \dots, x_n\}$, $|x| = N$

n-th subject's data $x_n = \{x_{n1}, x_{n2}, \dots, x_{nm}\}$, $|x_k| = M$

mean of population data $\mu = \frac{1}{N} \sum_N x_n$

mean of n-th subject data $\mu_n = \frac{1}{M} \sum_M x_{nm}$

standard deviation of population data $\sigma = \sqrt{\frac{1}{N} \sum_N (x_n - \mu)^2}$

standard deviation of n-th subject data $\sigma_n = \sqrt{\frac{1}{M} \sum_M (x_{nm} - \mu_n)^2}$

z-scored population data $^p x = \frac{x - \mu}{\sigma}$

n-th subject data z-scored using only data $^s x_n = \frac{x_n - \mu_n}{\sigma_n}$

N-th subject data, z-scored using population measures (μ, σ), and subjective measures afterwards ($^p \mu_n, ^p \sigma_n$, computed after first z-scoring):

$$^s p x_n = \frac{^p x_n - ^p \mu_n}{^p \sigma_n} = \frac{^p x_n - \frac{1}{M} \sum_M ^p x_{nm}}{\sqrt{\frac{1}{M} \sum_M (^p x_{nm} - \frac{1}{M} \sum_M ^p x_{nm})^2}}$$

By substituting $^p x_n$ with $\frac{x_n - \mu}{\sigma}$ we obtain

$$^s p x_n = \frac{\frac{x_n - \mu}{\sigma} - \frac{1}{M} \sum_M \frac{x_{nm} - \mu}{\sigma}}{\sqrt{\frac{1}{M} \sum_M (\frac{x_{nm} - \mu}{\sigma} - \frac{1}{M} \sum_M \frac{x_{nm} - \mu}{\sigma})^2}}$$

Which can be converted to

$$\begin{aligned} ^s p x_n &= \frac{Mx_n - M\mu - \frac{\sigma}{\sigma} \sum_M (x_{nm} - \mu)}{\sigma M \sqrt{\frac{1}{M} \sum_M (\frac{x_{nm} - \mu}{\sigma} - \frac{1}{\sigma M} \sum_M (x_{nm} - \mu))^2}} = \\ &= \frac{Mx_n - M\mu + M\mu - \sum_M x_{nm}}{M \sqrt{\frac{\sigma^2}{\sigma^2 M} \sum_M (x_{nm} - \mu - \frac{1}{M} \sum_M (x_{nm} - \mu))^2}} = \end{aligned}$$

$$= \frac{x_n - \frac{1}{M} \sum_M x_{nm}}{\sqrt{\frac{1}{M} \sum_M (x_{nm} - \mu + \mu - \frac{1}{M} \sum_M x_{nm})^2}} = \frac{x_n - \mu_n}{\sqrt{\frac{1}{M} \sum_M (x_{nm} - \mu_n)^2}} = \frac{x_n - \mu_n}{\sigma_n}$$

Thus, we obtain

$${}^{sp}x_n = \frac{x_n - \mu_n}{\sigma_n} = {}^s x_n$$

proving, that standardizing data twice results in the same transformation as z-scoring only once, using the latter method.

E.2 Personalized processing

E.2.1 Group models

Table E.1: Classification results in subject-dependent (group) experimental setup (metric \pm std).

Dataset	Model	Task	F1-macro	Precision-macro	Recall-macro
AMIGOS	Baseline (majority, test)	Arousal	0.38 \pm 0.02	0.31 \pm 0.02	0.50 \pm 0.00
		Valence	0.36 \pm 0.01	0.28 \pm 0.02	0.50 \pm 0.00
	Baseline (trained all)	Arousal	0.38 \pm 0.02	0.31 \pm 0.02	0.50 \pm 0.00
		Valence	0.36 \pm 0.01	0.28 \pm 0.02	0.50 \pm 0.00
AMIGOS	Initialized (trained all)	Arousal	0.62 \pm 0.04	0.63 \pm 0.04	0.62 \pm 0.04
		Valence	0.57 \pm 0.03	0.59 \pm 0.03	0.58 \pm 0.03
	Initialized (trained head)	Arousal	0.61 \pm 0.02	0.61 \pm 0.02	0.61 \pm 0.03
		Valence	0.56 \pm 0.03	0.57 \pm 0.03	0.56 \pm 0.02
ASCERTAIN	Baseline (majority, test)	Arousal	0.45 \pm 0.00	0.42 \pm 0.01	0.50 \pm 0.00
		Valence	0.38 \pm 0.01	0.31 \pm 0.01	0.50 \pm 0.00
	Baseline (trained all)	Arousal	0.45 \pm 0.00	0.42 \pm 0.01	0.50 \pm 0.00
		Valence	0.38 \pm 0.01	0.31 \pm 0.01	0.50 \pm 0.00
ASCERTAIN	Initialized (trained all)	Arousal	0.61 \pm 0.03	0.64 \pm 0.02	0.60 \pm 0.03
		Valence	0.55 \pm 0.02	0.55 \pm 0.02	0.55 \pm 0.02
	Initialized (trained head)	Arousal	0.51 \pm 0.01	0.62 \pm 0.09	0.52 \pm 0.01
		Valence	0.51 \pm 0.01	0.52 \pm 0.01	0.52 \pm 0.01
CASE	Baseline (majority, test)	Arousal	0.38 \pm 0.02	0.31 \pm 0.02	0.50 \pm 0.00
		Valence	0.39 \pm 0.01	0.33 \pm 0.01	0.50 \pm 0.00
	Baseline (trained all)	Arousal	0.45 \pm 0.08	0.43 \pm 0.12	0.52 \pm 0.02
		Valence	0.41 \pm 0.02	0.45 \pm 0.14	0.49 \pm 0.01
CASE	Initialized (trained all)	Arousal	0.57 \pm 0.04	0.59 \pm 0.02	0.59 \pm 0.03
		Valence	0.66 \pm 0.03	0.67 \pm 0.03	0.66 \pm 0.03
	Initialized (trained head)	Arousal	0.57 \pm 0.02	0.58 \pm 0.02	0.58 \pm 0.02
		Valence	0.56 \pm 0.02	0.59 \pm 0.04	0.56 \pm 0.02
DREAMER	Baseline (majority, test)	Arousal	0.42 \pm 0.02	0.36 \pm 0.02	0.50 \pm 0.00
		Valence	0.38 \pm 0.01	0.31 \pm 0.02	0.50 \pm 0.00
	Baseline (trained all)	Arousal	0.39 \pm 0.07	0.33 \pm 0.08	0.50 \pm 0.00
		Valence	0.35 \pm 0.04	0.28 \pm 0.05	0.50 \pm 0.00
DREAMER	Initialized (trained all)	Arousal	0.52 \pm 0.03	0.56 \pm 0.06	0.53 \pm 0.02
		Valence	0.55 \pm 0.09	0.62 \pm 0.05	0.59 \pm 0.06
	Initialized (trained head)	Arousal	0.50 \pm 0.07	0.54 \pm 0.04	0.53 \pm 0.03
		Valence	0.54 \pm 0.08	0.58 \pm 0.04	0.57 \pm 0.05

Table E.2: Regression results in subject-dependent (group) experimental setup (metric \pm std).

Dataset	Model	Task	RMSE	MAE	CCC
AMIGOS	Baseline (average, test)	Arousal	0.22 \pm 0.01	0.19 \pm 0.01	-0.00 \pm 0.00
		Valence	0.28 \pm 0.01	0.24 \pm 0.01	0.00 \pm 0.00
	Baseline (trained all)	Arousal	0.30 \pm 0.07	0.26 \pm 0.06	0.00 \pm 0.00
		Valence	0.53 \pm 0.26	0.48 \pm 0.26	0.00 \pm 0.00
	Initialized (trained all)	Arousal	0.32 \pm 0.02	0.26 \pm 0.01	0.29 \pm 0.08
		Valence	0.44 \pm 0.04	0.35 \pm 0.03	0.10 \pm 0.05
	Initialized (trained head)	Arousal	0.29 \pm 0.02	0.23 \pm 0.02	0.22 \pm 0.04
		Valence	0.46 \pm 0.03	0.38 \pm 0.03	0.11 \pm 0.03
ASCERTAIN	Baseline (average, test)	Arousal	0.23 \pm 0.01	0.18 \pm 0.01	-0.00 \pm 0.00
		Valence	0.30 \pm 0.01	0.26 \pm 0.01	0.00 \pm 0.00
	Baseline (trained all)	Arousal	0.52 \pm 0.24	0.48 \pm 0.25	0.00 \pm 0.00
		Valence	0.39 \pm 0.05	0.32 \pm 0.04	0.00 \pm 0.00
	Initialized (trained all)	Arousal	0.32 \pm 0.01	0.24 \pm 0.00	0.29 \pm 0.03
		Valence	0.46 \pm 0.02	0.36 \pm 0.01	0.14 \pm 0.03
	Initialized (trained head)	Arousal	0.31 \pm 0.02	0.25 \pm 0.02	0.19 \pm 0.03
		Valence	0.38 \pm 0.01	0.31 \pm 0.01	0.11 \pm 0.03
CASE	Baseline (average, test)	Arousal	0.18 \pm 0.01	0.14 \pm 0.01	0.00 \pm 0.00
		Valence	0.20 \pm 0.01	0.14 \pm 0.01	0.00 \pm 0.00
	Baseline (trained all)	Arousal	0.23 \pm 0.01	0.17 \pm 0.01	0.22 \pm 0.12
		Valence	0.70 \pm 0.56	0.65 \pm 0.59	0.13 \pm 0.11
	Initialized (trained all)	Arousal	0.21 \pm 0.02	0.16 \pm 0.02	0.42 \pm 0.06
		Valence	0.22 \pm 0.01	0.16 \pm 0.01	0.44 \pm 0.04
	Initialized (trained head)	Arousal	0.23 \pm 0.03	0.18 \pm 0.03	0.21 \pm 0.07
		Valence	0.35 \pm 0.10	0.29 \pm 0.10	0.07 \pm 0.07
DREAMER	Baseline (average, test)	Arousal	0.27 \pm 0.01	0.22 \pm 0.01	0.00 \pm 0.00
		Valence	0.33 \pm 0.01	0.28 \pm 0.02	0.00 \pm 0.00
	Baseline (trained all)	Arousal	0.57 \pm 0.19	0.52 \pm 0.19	0.00 \pm 0.00
		Valence	0.80 \pm 0.71	0.75 \pm 0.72	0.00 \pm 0.00
	Initialized (trained all)	Arousal	0.42 \pm 0.04	0.34 \pm 0.04	0.14 \pm 0.08
		Valence	0.50 \pm 0.03	0.41 \pm 0.03	0.15 \pm 0.11
	Initialized (trained head)	Arousal	0.48 \pm 0.09	0.37 \pm 0.10	0.08 \pm 0.07
		Valence	0.47 \pm 0.03	0.37 \pm 0.03	0.11 \pm 0.07

Table E.3: Results of Friedman’s test between processing methods in subject-dependent (group) experimental design, based on relative differences between models and baselines.

Dataset	Task	χ^2	p (χ^2)	F	p (F)	Signif. (χ^2 / F)
AMIGOS classification	arousal	20.67	0.000	19.11	0.000	*** / ***
	valence	7.81	0.167	1.82	0.155	ns / ns
AMIGOS regression	arousal	9.00	0.109	2.25	0.088	ns / ns
	valence	10.03	0.074	2.68	0.051	ns / ns
ASCERTAIN classification	arousal	13.80	0.016	4.93	0.004	* / **
	valence	13.57	0.018	4.75	0.005	* / **
ASCERTAIN regression	arousal	3.97	0.553	0.76	0.592	ns / ns
	valence	15.40	0.008	6.42	0.001	** / **
CASE classification	arousal	7.74	0.170	1.79	0.159	ns / ns
	valence	2.37	0.795	0.42	0.829	ns / ns
CASE regression	arousal	16.77	0.004	8.15	0.000	** / ***
	valence	6.14	0.292	1.30	0.302	ns / ns
DREAMER classification	arousal	2.94	0.708	0.53	0.748	ns / ns
	valence	3.79	0.579	0.72	0.619	ns / ns
DREAMER regression	arousal	1.69	0.890	0.29	0.913	ns / ns
	valence	3.63	0.604	0.68	0.644	ns / ns

Table E.4: Comparison of processing methods in subject-dependent (group) experimental design (metric±std). AMIGOS dataset, classification, majority baseline (test).

Task	Physiology standardization	Annotation standardization	F1-macro	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.383±0.018	-	-	-	-
	-	Within-person	0.443±0.031	-	-	-	-
Arousal	Not standardized	Not standardized	0.623±0.036	0.631±0.128	0.008	0.024	*
	Within-person	Within-person	0.558±0.051	0.271±0.196	0.016	0.048	*
	Within-person	Not standardized	0.634±0.036	0.663±0.148	0.008	0.024	*
	Within-person	Within-person	0.534±0.043	0.214±0.165	0.016	0.048	*
	Within-sample	Not standardized	0.631±0.037	0.656±0.157	0.008	0.024	*
	Within-person	Within-person	0.543±0.045	0.236±0.166	0.016	0.048	*
Valence baseline	-	Not standardized	0.356±0.013	-	-	-	-
	-	Within-person	0.402±0.055	-	-	-	-
Valence	Not standardized	Not standardized	0.568±0.033	0.594±0.056	0.008	0.024	*
	Within-person	Within-person	0.545±0.042	0.390±0.250	0.016	0.048	*
	Within-person	Not standardized	0.576±0.015	0.618±0.026	0.008	0.024	*
	Within-person	Within-person	0.555±0.032	0.414±0.237	0.016	0.048	*
	Within-sample	Not standardized	0.568±0.026	0.593±0.022	0.008	0.024	*
	Within-person	Within-person	0.539±0.024	0.371±0.220	0.016	0.048	*

Table E.5: Comparison of processing methods in subject-dependent (group) experimental design (metric±std). AMIGOS dataset, regression, average baseline (test).

Task	Physiology standardization	Annotation standardization	RMSE	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.223±0.009	-	-	-	-
	-	Within-person	0.167±0.030	-	-	-	-
Arousal	Not standardized	Not standardized	0.324±0.019	0.456±0.112	0.008	0.024	*
	Within-person	Within-person	0.395±0.133	1.564±1.279	0.008	0.024	*
	Within-person	Not standardized	0.292±0.023	0.307±0.079	0.008	0.024	*
	Within-person	Within-person	0.370±0.113	1.289±0.741	0.008	0.024	*
	Within-sample	Not standardized	0.309±0.022	0.391±0.131	0.008	0.024	*
	Within-person	Within-person	0.494±0.356	1.907±1.865	0.008	0.024	*
Valence baseline	-	Not standardized	0.279±0.012	-	-	-	-
	-	Within-person	0.174±0.033	-	-	-	-
Valence	Not standardized	Not standardized	0.438±0.040	0.574±0.145	0.008	0.024	*
	Within-person	Within-person	0.261±0.106	0.450±0.414	0.095	0.190	ns
	Within-person	Not standardized	0.462±0.101	0.654±0.354	0.008	0.024	*
	Within-person	Within-person	0.222±0.046	0.278±0.098	0.095	0.190	ns
	Within-sample	Not standardized	0.443±0.091	0.590±0.333	0.008	0.024	*
	Within-person	Within-person	0.253±0.041	0.512±0.369	0.032	0.095	ns

Table E.6: Results of Conover’s post-hoc test between processing methods in subject-dependent (group) experimental design (metric \pm std). AMIGOS dataset, arousal classification. Comparisons were done on Values adjusted for baselines (gains). P-values were adjusted using Holm’s procedure [209].

Physiology standardization (G1)	Annotations processing (G1)	Physiology standardization (G2)	Annotations processing (G2)	Statistic	p	Signif.
Not standardized	Not standardized	Within-sample	Not standardized	0.735	1.000	ns
	Within-person	Not standardized	Within-person	-4.779	0.001	**
	Within-person	Not standardized	Not standardized	4.044	0.004	**
	Within-sample	Within-sample	Not standardized	4.779	0.001	**
	Within-person	Within-person	Within-person	-0.735	1.000	ns
Within-person	Not standardized	Not standardized	Not standardized	-1.470	0.942	ns
	Within-person	Within-person	Within-person	-5.514	0.000	***
	Within-sample	Within-sample	Not standardized	-0.735	1.000	ns
	Within-person	Within-person	Within-person	-6.249	0.000	***
	Within-person	Not standardized	Not standardized	5.514	0.000	***
	Within-person	Within-person	Within-person	1.470	0.942	ns
	Within-person	Within-person	Not standardized	6.985	0.000	***
	Within-sample	Within-sample	Not standardized	6.249	0.000	***
	Within-person	Within-person	Within-person	0.735	1.000	ns
Within-sample	Within-person	Within-sample	Not standardized	5.514	0.000	***

Table E.7: Comparison of processing methods in subject-dependent (group) experimental design (metric \pm std). ASCERTAIN dataset, classification, majority baseline (test).

Task	Physiology standardization	Annotation standardization	F1-macro	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.455 \pm 0.004	-	-	-	-
	-	Within-person	0.427 \pm 0.011	-	-	-	-
Arousal	Not standardized	Not standardized	0.606 \pm 0.027	0.332 \pm 0.050	0.008	0.024	*
	Within-person	Within-person	0.533 \pm 0.026	0.247 \pm 0.048	0.008	0.024	*
	Within-person	Not standardized	0.615 \pm 0.025	0.351 \pm 0.045	0.008	0.024	*
	Within-person	Within-person	0.542 \pm 0.016	0.267 \pm 0.023	0.008	0.024	*
	Within-sample	Not standardized	0.606 \pm 0.019	0.331 \pm 0.033	0.008	0.024	*
	Within-person	Within-person	0.548 \pm 0.027	0.282 \pm 0.043	0.008	0.024	*
Valence baseline	-	Not standardized	0.379 \pm 0.007	-	-	-	-
	-	Within-person	0.395 \pm 0.018	-	-	-	-
Valence	Not standardized	Not standardized	0.547 \pm 0.015	0.442 \pm 0.060	0.008	0.024	*
	Within-person	Within-person	0.516 \pm 0.016	0.311 \pm 0.078	0.008	0.024	*
	Within-person	Not standardized	0.540 \pm 0.018	0.424 \pm 0.068	0.008	0.024	*
	Within-person	Within-person	0.532 \pm 0.023	0.349 \pm 0.064	0.008	0.024	*
	Within-sample	Not standardized	0.538 \pm 0.019	0.420 \pm 0.071	0.008	0.024	*
	Within-person	Within-person	0.531 \pm 0.020	0.347 \pm 0.048	0.008	0.024	*

Table E.8: Comparison of processing methods in subject-dependent (group) experimental design (metric \pm std). ASCERTAIN dataset, regression, average baseline (test).

Task	Physiology standardization	Annotation standardization	RMSE	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.229 \pm 0.009	-	-	-	-
	-	Within-person	0.158 \pm 0.005	-	-	-	-
Arousal	Not standardized	Not standardized	0.315 \pm 0.009	0.380 \pm 0.054	0.008	0.024	*
	Within-person	Within-person	0.305 \pm 0.103	0.925 \pm 0.612	0.008	0.024	*
	Within-person	Not standardized	0.314 \pm 0.006	0.373 \pm 0.035	0.008	0.024	*
	Within-person	Within-person	0.226 \pm 0.011	0.427 \pm 0.047	0.008	0.024	*
	Within-sample	Not standardized	0.309 \pm 0.005	0.354 \pm 0.068	0.008	0.024	*
	Within-person	Within-person	0.234 \pm 0.017	0.481 \pm 0.124	0.008	0.024	*
Valence baseline	-	Not standardized	0.296 \pm 0.006	-	-	-	-
	-	Within-person	0.180 \pm 0.006	-	-	-	-
Valence	Not standardized	Not standardized	0.464 \pm 0.019	0.569 \pm 0.084	0.008	0.024	*
	Within-person	Within-person	0.592 \pm 0.526	2.256 \pm 2.798	0.008	0.024	*
	Within-person	Not standardized	0.463 \pm 0.022	0.566 \pm 0.061	0.008	0.024	*
	Within-person	Within-person	0.531 \pm 0.351	1.938 \pm 1.894	0.008	0.024	*
	Within-sample	Not standardized	0.452 \pm 0.029	0.528 \pm 0.073	0.008	0.024	*
	Within-person	Within-person	0.431 \pm 0.083	1.392 \pm 0.430	0.008	0.024	*

Table E.9: Results of Conover’s post-hoc test between processing methods in subject-dependent (group) experimental design (metric±std). ASCERTAIN dataset, arousal classification. Comparisons were done on Values adjusted for baselines (gains). P-values were adjusted using Holm’s procedure [209].

Physiology standardization (G1)	Annotations processing (G1)	Physiology standardization (G2)	Annotations processing (G2)	Statistic	p	Signif.
Not standardized	Not standardized	Within-sample	Not standardized	-0.226	1.000	ns
	Within-person	Not standardized	Within-person	-1.581	1.000	ns
	Within-person	Not standardized	Not standardized	2.711	0.175	ns
	Within-sample	Within-sample	Not standardized	2.485	0.219	ns
	Within-person	Within-person	Within-person	1.129	1.000	ns
Within-person	Not standardized	Not standardized	Not standardized	-1.129	1.000	ns
	Within-person	Within-person	Within-person	-3.840	0.015	*
	Within-sample	Within-sample	Not standardized	-1.355	1.000	ns
	Within-person	Within-person	Within-person	-2.711	0.175	ns
	Within-person	Not standardized	Not standardized	2.711	0.175	ns
	Within-person	Within-person	Within-person	0.000	1.000	ns
	Within-person	Within-person	Not standardized	3.840	0.015	*
	Within-sample	Within-sample	Not standardized	2.485	0.219	ns
	Within-person	Within-person	Within-person	1.129	1.000	ns
Within-sample	Within-person	Within-sample	Not standardized	1.355	1.000	ns

Table E.10: Results of Conover’s post-hoc test between processing methods in subject-dependent (group) experimental design (metric±std). ASCERTAIN dataset, valence classification. Comparisons were done on Values adjusted for baselines (gains). P-values were adjusted using Holm’s procedure [209].

Physiology standardization (G1)	Annotations processing (G1)	Physiology standardization (G2)	Annotations processing (G2)	Statistic	p	Signif.
Not standardized	Not standardized	Within-sample	Not standardized	-1.118	1.000	ns
	Within-person	Not standardized	Within-person	-4.025	0.010	**
	Within-person	Not standardized	Not standardized	3.354	0.044	*
		Within-sample	Not standardized	2.236	0.369	ns
		Within-person	Within-person	-0.671	1.000	ns
Within-person	Not standardized	Not standardized	Not standardized	1.342	1.000	ns
		Within-person	Within-person	-2.012	0.521	ns
		Within-sample	Not standardized	0.224	1.000	ns
		Within-person	Within-person	-2.683	0.157	ns
	Within-person	Not standardized	Not standardized	2.907	0.113	ns
		Within-person	Within-person	-0.447	1.000	ns
		Within-person	Not standardized	1.565	0.932	ns
		Within-sample	Not standardized	1.789	0.710	ns
		Within-person	Within-person	-1.118	1.000	ns
Within-sample	Within-person	Within-sample	Not standardized	2.907	0.113	ns

Table E.11: Results of Conover's post-hoc test between processing methods in subject-dependent experimental design (metric \pm std). ASCERTAIN dataset, valence regression. Comparisons were done on Values adjusted for baselines (gains). P-values were adjusted using Holm's procedure [209].

Physiology processing (G1)	Annotations processing (G1)	Physiology processing (G2)	Annotations processing (G2)	Statistic	p	Signif.
Not standardized	Not standardized	Within-sample	Not standardized Within-person	-0.488 3.660	1.000 0.020	ns *
	Within-person	Not standardized	Not standardized	-2.440	0.169	ns
	Within-sample	Within-sample	Not standardized Within-person	-2.928 1.220	0.092 1.000	ns ns
Within-person	Not standardized	Not standardized	Not standardized Within-person	0.244 2.684	1.000 0.128	ns ns
	Within-sample	Within-sample	Not standardized Within-person	-0.244 3.904	1.000 0.012	ns *
	Within-person	Not standardized	Not standardized Within-person	-2.684 -0.244	0.128 1.000	ns ns
	Within-person	Within-person	Not standardized	-2.928	0.092	ns
	Within-sample	Within-sample	Not standardized Within-person	-3.172 0.976	0.058 1.000	ns ns
Within-sample	Within-person	Within-sample	Not standardized	-4.148	0.007	**

Table E.12: Comparison of processing methods in subject-dependent (group) experimental design (metric \pm std). CASE dataset, classification, majority baseline (test).

Task	Physiology standardization	Annotation standardization	F1-macro	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.379 \pm 0.018	-	-	-	-
	-	Within-person	0.408 \pm 0.026	-	-	-	-
Arousal	Not standardized	Not standardized	0.570 \pm 0.044	0.503 \pm 0.096	0.008	0.024	*
	Within-person	Within-person	0.686 \pm 0.024	0.689 \pm 0.136	0.008	0.024	*
	Not standardized	Not standardized	0.574 \pm 0.022	0.515 \pm 0.052	0.008	0.024	*
	Within-person	Within-person	0.701 \pm 0.047	0.725 \pm 0.188	0.008	0.024	*
	Not standardized	Not standardized	0.578 \pm 0.020	0.525 \pm 0.049	0.008	0.024	*
	Within-person	Within-person	0.691 \pm 0.037	0.701 \pm 0.165	0.008	0.024	*
Valence baseline	-	Not standardized	0.395 \pm 0.010	-	-	-	-
	-	Within-person	0.398 \pm 0.020	-	-	-	-
Valence	Not standardized	Not standardized	0.660 \pm 0.034	0.672 \pm 0.081	0.008	0.024	*
	Within-person	Within-person	0.656 \pm 0.088	0.647 \pm 0.220	0.008	0.024	*
	Not standardized	Not standardized	0.673 \pm 0.043	0.705 \pm 0.105	0.008	0.024	*
	Within-person	Within-person	0.651 \pm 0.055	0.637 \pm 0.160	0.008	0.024	*
	Not standardized	Not standardized	0.672 \pm 0.041	0.702 \pm 0.093	0.008	0.024	*
	Within-person	Within-person	0.641 \pm 0.061	0.612 \pm 0.162	0.008	0.024	*

Table E.13: Comparison of processing methods in subject-dependent (group) experimental design (metric \pm std). CASE dataset, regression, average baseline (test).

Task	Physiology standardization	Annotation standardization	RMSE	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.185 \pm 0.012	-	-	-	-
	-	Within-person	0.165 \pm 0.020	-	-	-	-
Arousal	Not standardized	Not standardized	0.206 \pm 0.019	0.116 \pm 0.098	0.095	0.190	ns
	Within-person	Within-person	0.169 \pm 0.029	0.019 \pm 0.093	0.841	1.000	ns
	Within-person	Not standardized	0.203 \pm 0.016	0.104 \pm 0.095	0.095	0.190	ns
	Within-person	Within-person	0.168 \pm 0.021	0.017 \pm 0.053	1.000	1.000	ns
	Within-sample	Not standardized	0.214 \pm 0.018	0.162 \pm 0.125	0.032	0.095	ns
	Within-person	Within-person	0.166 \pm 0.022	0.005 \pm 0.053	1.000	1.000	ns
Valence baseline	-	Not standardized	0.197 \pm 0.011	-	-	-	-
	-	Within-person	0.155 \pm 0.011	-	-	-	-
Valence	Not standardized	Not standardized	0.215 \pm 0.010	0.096 \pm 0.097	0.095	0.286	ns
	Within-person	Within-person	0.162 \pm 0.009	0.048 \pm 0.071	0.310	0.930	ns
	Within-person	Not standardized	0.199 \pm 0.013	0.011 \pm 0.080	0.841	1.000	ns
	Within-person	Within-person	0.154 \pm 0.021	-0.006 \pm 0.084	0.548	1.000	ns
	Within-sample	Not standardized	0.203 \pm 0.022	0.027 \pm 0.080	1.000	1.000	ns
	Within-person	Within-person	0.155 \pm 0.017	0.004 \pm 0.084	0.841	1.000	ns

Table E.14: Results of Conover’s post-hoc test between processing methods in subject-dependent experimental design (metric±std). CASE dataset, arousal regression. Comparisons were done on Values adjusted for baselines (gains). P-values were adjusted using Holm’s procedure [209].

Physiology processing (G1)	Annotations processing (G1)	Physiology processing (G2)	Annotations processing (G2)	Statistic	p	Signif.
Not standardized	Not standardized	Within-sample	Not standardized Within-person	0.000 -4.480	1.000 0.003	ns **
	Within-person	Not standardized	Not standardized	3.426	0.032	*
	Within-sample	Within-sample	Not standardized Within-person	3.426 -1.054	0.032 1.000	* ns
Within-person	Not standardized	Not standardized	Not standardized Within-person	0.527 -2.899	1.000 0.071	ns ns
	Within-sample	Within-sample	Not standardized Within-person	0.527 -3.953	1.000 0.010	ns *
	Within-person	Not standardized	Not standardized Within-person	3.426 0.000	0.032 1.000	* ns
	Within-person	Within-person	Not standardized	2.899	0.071	ns
	Within-sample	Within-sample	Not standardized Within-person	3.426 -1.054	0.032 1.000	* ns
Within-sample	Within-person	Within-sample	Not standardized	4.480	0.003	**

Table E.15: Comparison of processing methods in subject-dependent (group) experimental design (metric \pm std). DREAMER dataset, classification, majority baseline (test).

Task	Physiology standardization	Annotation standardization	F1-macro	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.420 \pm 0.016	-	-	-	-
	-	Within-person	0.379 \pm 0.020	-	-	-	-
Arousal	Not standardized	Not standardized	0.520 \pm 0.034	0.237 \pm 0.058	0.008	0.024	*
	Within-person	Within-person	0.511 \pm 0.049	0.358 \pm 0.183	0.008	0.024	*
	Within-person	Not standardized	0.527 \pm 0.033	0.256 \pm 0.098	0.008	0.024	*
	Within-person	Within-person	0.480 \pm 0.070	0.279 \pm 0.239	0.016	0.032	*
	Within-sample	Not standardized	0.534 \pm 0.033	0.275 \pm 0.105	0.008	0.024	*
	Within-person	Within-person	0.485 \pm 0.066	0.290 \pm 0.228	0.016	0.032	*
Valence baseline	-	Not standardized	0.379 \pm 0.013	-	-	-	-
	-	Within-person	0.353 \pm 0.011	-	-	-	-
Valence	Not standardized	Not standardized	0.552 \pm 0.092	0.456 \pm 0.231	0.008	0.024	*
	Within-person	Within-person	0.492 \pm 0.088	0.392 \pm 0.242	0.032	0.032	*
	Within-person	Not standardized	0.529 \pm 0.093	0.400 \pm 0.269	0.046	0.093	ns
	Within-person	Within-person	0.502 \pm 0.048	0.425 \pm 0.167	0.008	0.024	*
	Within-sample	Not standardized	0.500 \pm 0.091	0.324 \pm 0.253	0.046	0.093	ns
	Within-person	Within-person	0.502 \pm 0.060	0.426 \pm 0.194	0.008	0.024	*

Table E.16: Comparison of processing methods in subject-dependent (group) experimental design (metric \pm std). DREAMER dataset, regression, average baseline (test).

Task	Physiology standardization	Annotation standardization	RMSE	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.268 \pm 0.013	-	-	-	-
	-	Within-person	0.223 \pm 0.007	-	-	-	-
Arousal	Not standardized	Not standardized	0.421 \pm 0.041	0.577 \pm 0.200	0.008	0.024	*
	Within-person	Within-person	0.519 \pm 0.205	1.316 \pm 0.892	0.008	0.024	*
	Within-person	Not standardized	0.447 \pm 0.038	0.677 \pm 0.188	0.008	0.024	*
	Within-person	Within-person	0.385 \pm 0.061	0.736 \pm 0.317	0.008	0.024	*
	Within-sample	Not standardized	0.422 \pm 0.020	0.578 \pm 0.110	0.008	0.024	*
	Within-person	Within-person	0.399 \pm 0.065	0.798 \pm 0.329	0.008	0.024	*
Valence baseline	-	Not standardized	0.328 \pm 0.014	-	-	-	-
	-	Within-person	0.241 \pm 0.014	-	-	-	-
Valence	Not standardized	Not standardized	0.502 \pm 0.033	0.534 \pm 0.099	0.008	0.024	*
	Within-person	Within-person	0.501 \pm 0.203	1.091 \pm 0.843	0.008	0.024	*
	Within-person	Not standardized	0.517 \pm 0.079	0.570 \pm 0.165	0.008	0.024	*
	Within-person	Within-person	0.365 \pm 0.053	0.520 \pm 0.229	0.008	0.024	*
	Within-sample	Not standardized	0.511 \pm 0.096	0.548 \pm 0.214	0.008	0.024	*
	Within-person	Within-person	0.400 \pm 0.066	0.669 \pm 0.291	0.008	0.024	*

E.2.2 Subject-specific models

Table E.17: Classification results in subject-dependent (subject) experimental setup (metric \pm std).

Dataset	Model	Task	F1-macro	Precision-macro	Recall-macro
AMIGOS	Baseline (majority, test)	Arousal	0.37 \pm 0.18	0.33 \pm 0.18	0.44 \pm 0.18
		Valence	0.42 \pm 0.12	0.37 \pm 0.13	0.53 \pm 0.11
	Initialized (trained all)	Arousal	0.74 \pm 0.11	0.77 \pm 0.10	0.76 \pm 0.11
		Valence	0.70 \pm 0.11	0.73 \pm 0.11	0.73 \pm 0.10
	Initialized (trained head)	Arousal	0.79 \pm 0.08	0.81 \pm 0.08	0.81 \pm 0.07
		Valence	0.70 \pm 0.12	0.73 \pm 0.12	0.73 \pm 0.11
ASCERTAIN	Baseline (majority, test)	Arousal	0.39 \pm 0.09	0.34 \pm 0.09	0.46 \pm 0.09
		Valence	0.38 \pm 0.03	0.31 \pm 0.03	0.49 \pm 0.04
	Initialized (trained all)	Arousal	0.61 \pm 0.10	0.65 \pm 0.09	0.65 \pm 0.08
		Valence	0.55 \pm 0.07	0.59 \pm 0.08	0.60 \pm 0.06
	Initialized (trained head)	Arousal	0.60 \pm 0.09	0.65 \pm 0.08	0.64 \pm 0.07
		Valence	0.56 \pm 0.07	0.62 \pm 0.07	0.60 \pm 0.06
CASE	Baseline (majority, test)	Arousal	0.62 \pm 0.08	0.58 \pm 0.09	0.68 \pm 0.07
		Valence	0.32 \pm 0.19	0.29 \pm 0.20	0.38 \pm 0.19
	Initialized (trained all)	Arousal	0.72 \pm 0.09	0.75 \pm 0.10	0.76 \pm 0.08
		Valence	0.76 \pm 0.11	0.78 \pm 0.12	0.78 \pm 0.10
	Initialized (trained head)	Arousal	0.72 \pm 0.08	0.76 \pm 0.09	0.75 \pm 0.07
		Valence	0.76 \pm 0.12	0.78 \pm 0.13	0.79 \pm 0.11
DREAMER	Baseline (majority, test)	Arousal	0.34 \pm 0.07	0.29 \pm 0.07	0.41 \pm 0.07
		Valence	0.41 \pm 0.12	0.37 \pm 0.12	0.48 \pm 0.12
	Initialized (trained all)	Arousal	0.71 \pm 0.09	0.74 \pm 0.09	0.74 \pm 0.08
		Valence	0.75 \pm 0.09	0.77 \pm 0.09	0.78 \pm 0.08
	Initialized (trained head)	Arousal	0.72 \pm 0.10	0.74 \pm 0.11	0.76 \pm 0.09
		Valence	0.74 \pm 0.10	0.77 \pm 0.10	0.78 \pm 0.08

Table E.18: Regression results in subject-dependent (subject) experimental setup (metric \pm std).

Dataset	Model	Task	RMSE	MAE	CCC
AMIGOS	Baseline (average, test)	Arousal	0.20 \pm 0.07	0.19 \pm 0.07	0.00 \pm 0.00
		Valence	0.25 \pm 0.08	0.24 \pm 0.08	0.00 \pm 0.00
	Initialized (trained all)	Arousal	0.54 \pm 0.25	0.49 \pm 0.25	0.23 \pm 0.17
		Valence	0.51 \pm 0.21	0.46 \pm 0.20	0.27 \pm 0.18
	Initialized (trained head)	Arousal	0.62 \pm 0.29	0.58 \pm 0.30	0.18 \pm 0.14
		Valence	0.56 \pm 0.27	0.52 \pm 0.27	0.26 \pm 0.17
ASCERTAIN	Baseline (average, test)	Arousal	0.24 \pm 0.04	0.20 \pm 0.03	0.00 \pm 0.00
		Valence	0.28 \pm 0.04	0.25 \pm 0.04	0.00 \pm 0.00
	Initialized (trained all)	Arousal	0.48 \pm 0.18	0.41 \pm 0.18	0.23 \pm 0.14
		Valence	0.53 \pm 0.16	0.45 \pm 0.14	0.17 \pm 0.10
	Initialized (trained head)	Arousal	0.49 \pm 0.17	0.42 \pm 0.17	0.21 \pm 0.12
		Valence	0.55 \pm 0.16	0.47 \pm 0.16	0.17 \pm 0.10
CASE	Baseline (average, test)	Arousal	0.14 \pm 0.02	0.12 \pm 0.02	0.00 \pm 0.00
		Valence	0.15 \pm 0.03	0.13 \pm 0.03	0.00 \pm 0.00
	Initialized (trained all)	Arousal	0.29 \pm 0.18	0.25 \pm 0.18	0.29 \pm 0.12
		Valence	0.36 \pm 0.22	0.32 \pm 0.21	0.28 \pm 0.15
	Initialized (trained head)	Arousal	0.34 \pm 0.17	0.30 \pm 0.17	0.22 \pm 0.11
		Valence	0.39 \pm 0.20	0.35 \pm 0.20	0.22 \pm 0.11
DREAMER	Baseline (average, test)	Arousal	0.24 \pm 0.05	0.21 \pm 0.05	0.00 \pm 0.00
		Valence	0.24 \pm 0.04	0.22 \pm 0.03	0.00 \pm 0.00
	Initialized (trained all)	Arousal	0.48 \pm 0.18	0.43 \pm 0.17	0.32 \pm 0.14
		Valence	0.51 \pm 0.17	0.45 \pm 0.15	0.31 \pm 0.12
	Initialized (trained head)	Arousal	0.70 \pm 0.30	0.64 \pm 0.30	0.21 \pm 0.15
		Valence	0.57 \pm 0.24	0.52 \pm 0.23	0.30 \pm 0.14

Table E.19: Results of Friedman test between processing methods in subject-dependent (subject) experimental design, based on relative differences between models and baselines.

Dataset	Task	χ^2	p (χ^2)	F	p (F)	Signif. (χ^2 / F)
AMIGOS classification	arousal	0.31	0.958	0.10	0.959	ns / ns
	valence	5.82	0.120	1.99	0.119	ns / ns
AMIGOS regression	arousal	16.48	0.000	6.25	0.000	*** / ***
	valence	9.62	0.022	3.41	0.020	* / *
ASCERTAIN classification	arousal	14.22	0.002	5.13	0.002	** / **
	valence	36.32	0.000	15.57	0.000	*** / ***
ASCERTAIN regression	arousal	8.01	0.045	2.76	0.044	* / *
	valence	1.84	0.606	0.61	0.610	ns / ns
CASE classification	arousal	29.28	0.000	13.98	0.000	*** / ***
	valence	1.24	0.743	0.41	0.749	ns / ns
CASE regression	arousal	15.08	0.001	5.84	0.001	** / **
	valence	6.16	0.104	2.13	0.102	ns / ns
DREAMER classification	arousal	10.24	0.016	3.83	0.013	* / *
	valence	1.31	0.726	0.43	0.734	ns / ns
DREAMER regression	arousal	9.05	0.028	3.32	0.024	* / *
	valence	1.19	0.754	0.39	0.762	ns / ns

Table E.20: Comparison of processing methods in subject-dependent (subject) experimental design (metric \pm std). AMIGOS dataset, classification, majority baseline (test).

Task	Physiology standardization	Annotation standardization	F1-macro	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.331 \pm 0.254	-	-	-	-
	-	Within-person	0.369 \pm 0.177	-	-	-	-
Arousal	Within-person	Not standardized	0.806 \pm 0.145	2.446 \pm 3.199	0.000	0.000	***
		Within-person	0.743 \pm 0.109	1.958 \pm 2.717	0.000	0.000	***
Valence baseline	Within-sample	Not standardized	0.776 \pm 0.150	2.315 \pm 3.188	0.000	0.000	***
		Within-person	0.759 \pm 0.091	2.035 \pm 2.769	0.000	0.000	***
Valence	Within-person	Not standardized	0.368 \pm 0.125	-	-	-	-
		Within-person	0.424 \pm 0.122	-	-	-	-
Valence	Within-sample	Not standardized	0.721 \pm 0.140	1.065 \pm 0.817	0.000	0.000	***
		Within-person	0.695 \pm 0.108	0.745 \pm 0.517	0.000	0.000	***
Valence	Within-person	Not standardized	0.691 \pm 0.135	0.967 \pm 0.727	0.000	0.000	***
		Within-person	0.709 \pm 0.125	0.769 \pm 0.512	0.000	0.000	***

Table E.21: Comparison of processing methods in subject-dependent (subject) experimental design (metric \pm std). AMIGOS dataset, regression, average baseline (test).

Task	Physiology standardization	Annotation standardization	RMSE	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.118 \pm 0.057	-	-	-	-
	-	Within-person	0.196 \pm 0.068	-	-	-	-
Arousal	Within-person	Not standardized	0.413 \pm 0.258	4.126 \pm 6.558	0.000	0.000	***
		Within-person	0.537 \pm 0.247	2.498 \pm 4.221	0.000	0.000	***
Valence baseline	-	Not standardized	0.403 \pm 0.193	4.527 \pm 9.020	0.000	0.000	***
		Within-person	0.543 \pm 0.204	2.679 \pm 5.130	0.000	0.000	***
Valence	Within-person	Not standardized	0.200 \pm 0.087	-	-	-	-
		Within-person	0.252 \pm 0.081	-	-	-	-
Valence	Within-person	Not standardized	0.475 \pm 0.205	1.952 \pm 2.820	0.000	0.000	***
		Within-person	0.509 \pm 0.207	1.260 \pm 1.698	0.000	0.000	***
Valence	Within-sample	Not standardized	0.450 \pm 0.195	1.658 \pm 2.155	0.000	0.000	***
		Within-person	0.482 \pm 0.186	1.025 \pm 1.102	0.000	0.000	***

Table E.22: Results of Conover's post-hoc test between processing methods in subject-dependent (subject) experimental design (metric±std). AMIGOS dataset, arousal regression. Comparisons were done on Values adjusted for baselines (gains). P-values were adjusted using Holm's procedure [209].

Physiology standardization (G1)	Annotations processing (G1)	Physiology standardization (G2)	Annotations processing (G2)	Statistic	p	Signif.
Within-person	Not standardized	Within-sample	Not standardized	0.190	1.000	ns
	Within-person	Within-person	Within-person	-3.223	0.008	**
	Within-person	Within-person	Not standardized	2.654	0.027	*
	Within-person	Within-sample	Not standardized	2.844	0.021	*
Within-sample	Within-person	Within-sample	Within-person	-0.569	1.000	ns
	Within-person	Within-sample	Not standardized	3.413	0.005	**

Table E.23: Results of Conover’s post-hoc test between processing methods in subject-dependent (subject) experimental design (metric±std). AMIGOS dataset, valence regression. Comparisons were done on Values adjusted for baselines (gains). P-values were adjusted using Holm’s procedure [209].

Physiology standardization (G1)	Annotations processing (G1)	Physiology standardization (G2)	Annotations processing (G2)	Statistic	p	Signif.
Within-person	Not standardized	Within-sample	Not standardized	-1.239	0.654	ns
			Within-person	-2.890	0.028	*
	Within-person	Within-person	Not standardized	2.477	0.074	ns
		Within-sample	Not standardized	1.239	0.654	ns
			Within-person	-0.413	0.681	ns
Within-sample	Within-person	Within-sample	Not standardized	1.651	0.406	ns

Table E.24: Comparison of processing methods in subject-dependent (subject) experimental design (metric±std). ASCERTAIN dataset, classification, majority baseline (test).

Task	Physiology standardization	Annotation standardization	F1-macro	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.291±0.126	-	-	-	-
	-	Within-person	0.393±0.091	-	-	-	-
Arousal	Within-person	Not standardized	0.724±0.127	1.771±2.084	0.000	0.000	***
		Within-person	0.610±0.100	0.649±0.557	0.000	0.000	***
Valence baseline	-	Not standardized	0.718±0.127	1.756±2.139	0.000	0.000	***
		Within-person	0.606±0.093	0.637±0.530	0.000	0.000	***
Valence	Within-person	Not standardized	0.378±0.037	-	-	-	-
		Within-person	0.379±0.026	-	-	-	-
Valence	Within-person	Not standardized	0.612±0.081	0.664±0.512	0.000	0.000	***
		Within-person	0.554±0.071	0.472±0.240	0.000	0.000	***
Valence	Within-sample	Not standardized	0.607±0.075	0.648±0.493	0.000	0.000	***
		Within-person	0.601±0.068	0.601±0.269	0.000	0.000	***

Table E.25: Comparison of processing methods in subject-dependent (subject) experimental design (metric±std). ASCERTAIN dataset, regression, average baseline (test).

Task	Physiology standardization	Annotation standardization	RMSE	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.189±0.048	-	-	-	-
	-	Within-person	0.237±0.036	-	-	-	-
Arousal	Within-person	Not standardized	0.405±0.193	1.217±1.043	0.000	0.000	***
		Within-person	0.477±0.180	1.024±0.714	0.000	0.000	***
Valence baseline	Within-sample	Not standardized	0.403±0.155	1.245±1.060	0.000	0.000	***
		Within-person	0.456±0.160	0.951±0.727	0.000	0.000	***
Valence	Within-person	Not standardized	0.271±0.047	-	-	-	-
		Within-person	0.282±0.042	-	-	-	-
Valence	Within-sample	Not standardized	0.493±0.136	0.851±0.553	0.000	0.000	***
		Within-person	0.533±0.156	0.901±0.531	0.000	0.000	***
Valence	Within-person	Not standardized	0.499±0.139	0.872±0.547	0.000	0.000	***
		Within-person	0.512±0.145	0.837±0.521	0.000	0.000	***

Table E.26: Results of Conover’s post-hoc test between processing methods in subject-dependent (subject) experimental design (metric±std). ASCERTAIN dataset, arousal classification. Comparisons were done on Values adjusted for baselines (gains). P-values were adjusted using Holm’s procedure [209].

Physiology standardization (G1)	Annotations processing (G1)	Physiology standardization (G2)	Annotations processing (G2)	Statistic	p	Signif.
Within-person	Not standardized	Within-sample	Not standardized Within-person	-0.642 -3.171	1.000 0.011	ns *
Within-person	Within-person	Within-person	Not standardized	2.931	0.020	*
Within-sample	Within-sample	Within-sample	Not standardized Within-person	2.288 -0.241	0.071 1.000	ns ns
Within-sample	Within-person	Within-sample	Not standardized	2.529	0.050	*

Table E.27: Results of Conover’s post-hoc test between processing methods in subject-dependent (subject) experimental design (metric±std). ASCERTAIN dataset, valence classification. Comparisons were done on Values adjusted for baselines (gains). P-values were adjusted using Holm’s procedure [209].

Physiology standardization (G1)	Annotations processing (G1)	Physiology standardization (G2)	Annotations processing (G2)	Statistic	p	Signif.
Within-person	Not standardized	Within-sample	Not standardized Within-person	-1.001 -0.218	0.955 0.955	ns ns
Within-person	Within-person	Within-person	Not standardized	5.919	0.000	***
Within-sample	Within-sample	Within-sample	Not standardized Within-person	4.918 5.702	0.000 0.000	*** ***
Within-sample	Within-person	Within-sample	Not standardized	-0.783	0.955	ns

Table E.28: Results of Conover’s post-hoc test between processing methods in subject-dependent (subject) experimental design (metric±std). ASCERTAIN dataset, arousal regression. Comparisons were done on Values adjusted for baselines (gains). P-values were adjusted using Holm’s procedure [209].

Physiology standardization (G1)	Annotations processing (G1)	Physiology standardization (G2)	Annotations processing (G2)	Statistic	p	Signif.
Within-person	Not standardized	Within-sample	Not standardized	-0.234	0.815	ns
			Within-person	-2.579	0.065	ns
	Within-person	Within-person	Not standardized	1.251	0.744	ns
		Within-sample	Not standardized	1.016	0.744	ns
			Within-person	-1.329	0.744	ns
Within-sample	Within-person	Within-sample	Not standardized	2.345	0.102	ns

Table E.29: Comparison of processing methods in subject-dependent (subject) experimental design (metric \pm std). CASE dataset, classification, majority baseline (test).

Task	Physiology standardization	Annotation standardization	F1-macro	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.483 \pm 0.122	-	-	-	-
	-	Within-person	0.618 \pm 0.080	-	-	-	-
Arousal	Within-person	Not standardized	0.700 \pm 0.103	0.613 \pm 0.759	0.000	0.000	***
	Within-person	Within-person	0.723 \pm 0.094	0.181 \pm 0.172	0.000	0.000	***
Valence baseline	Within-sample	Not standardized	0.701 \pm 0.118	0.612 \pm 0.757	0.000	0.000	***
	Within-sample	Within-person	0.728 \pm 0.096	0.187 \pm 0.147	0.000	0.000	***
Valence	Within-person	Not standardized	0.282 \pm 0.097	-	-	-	-
	Within-person	Within-person	0.323 \pm 0.194	-	-	-	-
Valence	Within-person	Not standardized	0.742 \pm 0.112	2.175 \pm 1.882	0.000	0.000	***
	Within-sample	Within-person	0.755 \pm 0.114	2.305 \pm 1.884	0.000	0.000	***
Valence	Within-person	Not standardized	0.761 \pm 0.111	2.253 \pm 1.917	0.000	0.000	***
	Within-sample	Within-person	0.743 \pm 0.126	2.262 \pm 1.875	0.000	0.000	***

Table E.30: Comparison of processing methods in subject-dependent (subject) experimental design (metric \pm std). CASE dataset, regression, average baseline (test).

Task	Physiology standardization	Annotation standardization	RMSE	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.100 \pm 0.029	-	-	-	-
	-	Within-person	0.142 \pm 0.021	-	-	-	-
Arousal	Within-person	Not standardized	0.303 \pm 0.233	2.516 \pm 3.862	0.000	0.000	***
		Within-person	0.285 \pm 0.184	1.064 \pm 1.434	0.000	0.000	***
Valence baseline	Within-sample	Not standardized	0.277 \pm 0.163	2.573 \pm 4.282	0.000	0.000	***
		Within-person	0.305 \pm 0.173	1.179 \pm 1.217	0.000	0.000	***
Valence	Within-person	Not standardized	0.107 \pm 0.041	-	-	-	-
		Within-person	0.148 \pm 0.026	-	-	-	-
Valence	Within-sample	Not standardized	0.344 \pm 0.241	3.344 \pm 4.691	0.000	0.000	***
		Within-person	0.360 \pm 0.218	1.497 \pm 1.553	0.000	0.000	***
Valence	Within-person	Not standardized	0.299 \pm 0.131	2.692 \pm 3.354	0.000	0.000	***
		Within-person	0.344 \pm 0.159	1.392 \pm 1.181	0.000	0.000	***

Table E.31: Results of Conover's post-hoc test between processing methods in subject-dependent (subject) experimental design (metric±std). CASE dataset, arousal classification. Comparisons were done on Values adjusted for baselines (gains). P-values were adjusted using Holm's procedure [209].

Physiology standardization (G1)	Annotations processing (G1)	Physiology standardization (G2)	Annotations processing (G2)	Statistic	p	Signif.
Within-person	Not standardized	Within-sample	Not standardized Within-person	0.239 -4.788	0.949 0.000	ns ***
	Within-person	Within-person	Not standardized	4.070	0.000	***
		Within-sample	Not standardized Within-person	4.309 -0.718	0.000 0.949	*** ns
Within-sample	Within-person	Within-sample	Not standardized	5.027	0.000	***

Table E.32: Results of Conover's post-hoc test between processing methods in subject-dependent (subject) experimental design (metric±std). CASE dataset, arousal regression. Comparisons were done on Values adjusted for baselines (gains). P-values were adjusted using Holm's procedure [209].

Physiology standardization (G1)	Annotations processing (G1)	Physiology standardization (G2)	Annotations processing (G2)	Statistic	p	Signif.
Within-person	Not standardized	Within-sample	Not standardized Within-person	-0.108 -2.586	0.914 0.045	ns *
	Within-person	Within-person	Not standardized	3.341	0.007	**
		Within-sample	Not standardized Within-person	3.233 0.754	0.009 0.905	** ns
Within-sample	Within-person	Within-sample	Not standardized	2.478	0.045	*

Table E.33: Comparison of processing methods in subject-dependent (subject) experimental design (metric \pm std). DREAMER dataset, classification, majority baseline (test).

Task	Physiology standardization	Annotation standardization	F1-macro	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.257 \pm 0.102	-	-	-	-
	-	Within-person	0.340 \pm 0.068	-	-	-	-
Arousal	Within-person	Not standardized	0.751 \pm 0.104	2.834 \pm 2.702	0.000	0.000	***
		Within-person	0.707 \pm 0.088	1.167 \pm 0.519	0.000	0.000	***
Valence baseline	-	Not standardized	0.777 \pm 0.114	2.904 \pm 2.594	0.000	0.000	***
		Within-person	0.729 \pm 0.083	1.254 \pm 0.619	0.000	0.000	***
Valence	Within-person	Not standardized	0.400 \pm 0.110	-	-	-	-
		Within-person	0.412 \pm 0.116	-	-	-	-
Valence	Within-person	Not standardized	0.732 \pm 0.080	1.000 \pm 0.683	0.000	0.000	***
		Within-person	0.747 \pm 0.090	0.977 \pm 0.658	0.000	0.000	***
Valence	Within-sample	Not standardized	0.743 \pm 0.073	1.016 \pm 0.647	0.000	0.000	***
		Within-person	0.747 \pm 0.080	0.977 \pm 0.636	0.000	0.000	***

Table E.34: Comparison of processing methods in subject-dependent (subject) experimental design (metric \pm std). DREAMER dataset, regression, average baseline (test).

Task	Physiology standardization	Annotation standardization	RMSE	Rel. diff. (gain)	p	p (adj.)	Signif.
Arousal baseline	-	Not standardized	0.197 \pm 0.045	-	-	-	-
	-	Within-person	0.239 \pm 0.052	-	-	-	-
Arousal	Within-person	Not standardized	0.481 \pm 0.239	1.596 \pm 1.447	0.000	0.000	***
		Within-person	0.482 \pm 0.178	1.081 \pm 0.822	0.000	0.000	***
Valence baseline	Within-sample	Not standardized	0.448 \pm 0.155	1.423 \pm 1.003	0.000	0.000	***
		Within-person	0.500 \pm 0.216	1.120 \pm 0.798	0.000	0.000	***
Valence	Within-person	Not standardized	0.244 \pm 0.037	-	-	-	-
		Within-person	0.244 \pm 0.037	-	-	-	-
Valence	Within-sample	Not standardized	0.506 \pm 0.166	1.105 \pm 0.733	0.000	0.000	***
		Within-person	0.506 \pm 0.166	1.106 \pm 0.733	0.000	0.000	***
Valence	Within-person	Not standardized	0.516 \pm 0.159	1.138 \pm 0.651	0.000	0.000	***
		Within-person	0.515 \pm 0.159	1.138 \pm 0.651	0.000	0.000	***

Table E.35: Results of Conover's post-hoc test between processing methods in subject-dependent (subject) experimental design (metric±std). DREAMER dataset, arousal classification. Comparisons were done on Values adjusted for baselines (gains). P-values were adjusted using Holm's procedure [209].

Physiology standardization (G1)	Annotations processing (G1)	Physiology standardization (G2)	Annotations processing (G2)	Statistic	p	Signif.
Within-person	Not standardized	Within-sample	Not standardized	0.731	0.935	ns
	Within-person	Within-person	Within-person	-0.426	0.935	ns
	Within-person	Within-person	Not standardized	2.498	0.075	ns
		Within-sample	Not standardized	3.229	0.012	*
		Within-person	Within-person	2.071	0.169	ns
Within-sample	Within-person	Within-sample	Not standardized	1.157	0.754	ns

Table E.36: Results of Conover’s post-hoc test between processing methods in subject-dependent (subject) experimental design (metric±std). DREAMER dataset, arousal regression. Comparisons were done on Values adjusted for baselines (gains). P-values were adjusted using Holm’s procedure [209].

Physiology standardization (G1)	Annotations processing (G1)	Physiology standardization (G2)	Annotations processing (G2)	Statistic	p	Signif.
Within-person	Not standardized	Within-sample	Not standardized	-0.241	0.945	ns
	Within-person	Within-person	Within-person	-1.926	0.234	ns
			Not standardized	2.648	0.061	ns
		Within-sample	Not standardized	2.407	0.094	ns
			Within-person	0.722	0.945	ns
Within-sample	Within-person	Within-sample	Not standardized	1.685	0.290	ns

E.2.3 Design comparisons

Table E.37: Comparison of results in experimental setups (designs) with respective baselines (metric \pm std). Arousal classification, majority baseline (test). Note - gain represents scores relative to the baseline, averaged over a set of participants (may differ from simple calculation based on average values).

Dataset	Setup	Predictor	F1-macro	Gain	p	Signif.
AMIGOS	Subject-independent	Baseline	0.39 \pm 0.19	-	-	-
		Model	0.43 \pm 0.20	0.04 \pm 0.45	0.931	ns
	Subject-dependent (group)	Baseline	0.32 \pm 0.25	-	-	-
		Model	0.57 \pm 0.25	1.15 \pm 1.97	0.000	***
	Subject-dependent (subject)	Baseline	0.37 \pm 0.18	-	-	-
		Model	0.74 \pm 0.11	1.96 \pm 2.72	0.000	***
ASCERTAIN	Subject-independent	Baseline	0.42 \pm 0.13	-	-	-
		Model	0.48 \pm 0.12	0.09 \pm 0.29	0.073	ns
	Subject-dependent (group)	Baseline	0.28 \pm 0.12	-	-	-
		Model	0.61 \pm 0.16	1.31 \pm 1.61	0.000	***
	Subject-dependent (subject)	Baseline	0.39 \pm 0.09	-	-	-
		Model	0.61 \pm 0.10	0.65 \pm 0.56	0.000	***
CASE	Subject-independent	Baseline	0.39 \pm 0.04	-	-	-
		Model	0.43 \pm 0.12	0.10 \pm 0.34	0.157	ns
	Subject-dependent (group)	Baseline	0.48 \pm 0.12	-	-	-
		Model	0.44 \pm 0.14	0.02 \pm 0.65	0.119	ns
	Subject-dependent (subject)	Baseline	0.62 \pm 0.08	-	-	-
		Model	0.72 \pm 0.09	0.18 \pm 0.17	0.000	***
DREAMER	Subject-independent	Baseline	0.42 \pm 0.05	-	-	-
		Model	0.47 \pm 0.09	0.15 \pm 0.26	0.017	*
	Subject-dependent (group)	Baseline	0.25 \pm 0.09	-	-	-
		Model	0.49 \pm 0.18	1.77 \pm 2.77	0.000	***
	Subject-dependent (subject)	Baseline	0.34 \pm 0.07	-	-	-
		Model	0.71 \pm 0.09	1.17 \pm 0.52	0.000	***

Table E.38: Comparison of results in experimental setups (designs) with respective baselines (metric \pm std). Valence classification, majority baseline (test). Note - gain represents scores relative to the baseline, averaged over a set of participants (may differ from simple calculation based on average values).

Dataset	Setup	Predictor	F1-macro	Gain	p	Signif.
AMIGOS	Subject-independent	Baseline	0.37 \pm 0.07	-	-	-
		Model	0.46 \pm 0.12	0.25 \pm 0.33	0.000	***
	Subject-dependent (group)	Baseline	0.46 \pm 0.13	-	-	-
		Model	0.47 \pm 0.18	0.12 \pm 0.73	0.546	ns
	Subject-dependent (subject)	Baseline	0.42 \pm 0.12	-	-	-
		Model	0.70 \pm 0.11	0.74 \pm 0.52	0.000	***
ASCERTAIN	Subject-independent	Baseline	0.38 \pm 0.03	-	-	-
		Model	0.44 \pm 0.10	0.17 \pm 0.28	0.000	***
	Subject-dependent (group)	Baseline	0.38 \pm 0.04	-	-	-
		Model	0.44 \pm 0.09	0.22 \pm 0.48	0.000	***
	Subject-dependent (subject)	Baseline	0.38 \pm 0.03	-	-	-
		Model	0.55 \pm 0.07	0.47 \pm 0.24	0.000	***
CASE	Subject-independent	Baseline	0.40 \pm 0.03	-	-	-
		Model	0.63 \pm 0.19	0.59 \pm 0.48	0.000	***
	Subject-dependent (group)	Baseline	0.28 \pm 0.10	-	-	-
		Model	0.61 \pm 0.17	1.68 \pm 1.85	0.000	***
	Subject-dependent (subject)	Baseline	0.32 \pm 0.19	-	-	-
		Model	0.76 \pm 0.11	2.31 \pm 1.88	0.000	***
DREAMER	Subject-independent	Baseline	0.39 \pm 0.04	-	-	-
		Model	0.49 \pm 0.12	0.26 \pm 0.33	0.004	**
	Subject-dependent (group)	Baseline	0.42 \pm 0.11	-	-	-
		Model	0.46 \pm 0.12	0.18 \pm 0.47	0.520	ns
	Subject-dependent (subject)	Baseline	0.41 \pm 0.12	-	-	-
		Model	0.75 \pm 0.09	0.98 \pm 0.66	0.000	***

Table E.39: Comparison of results in experimental setups (designs) with respective baselines (metric \pm std). Arousal regression, average baseline (test). Note - gain represents scores relative to the baseline, averaged over a set of participants (may differ from simple calculation based on average values).

Dataset	Setup	Predictor	RMSE	Gain	p	Signif.
AMIGOS	Subject-independent	Baseline	0.18 \pm 0.06	-	-	-
		Model	0.54 \pm 0.34	2.39 \pm 2.48	0.000	***
	Subject-dependent (group)	Baseline	0.10 \pm 0.04	-	-	-
		Model	0.28 \pm 0.09	2.29 \pm 2.44	0.000	***
	Subject-dependent (subject)	Baseline	0.20 \pm 0.07	-	-	-
		Model	0.54 \pm 0.25	2.46 \pm 4.17	0.000	***
ASCERTAIN	Subject-independent	Baseline	0.21 \pm 0.05	-	-	-
		Model	0.47 \pm 0.29	1.38 \pm 1.51	0.000	***
	Subject-dependent (group)	Baseline	0.19 \pm 0.05	-	-	-
		Model	0.29 \pm 0.10	0.55 \pm 0.41	0.000	***
	Subject-dependent (subject)	Baseline	0.24 \pm 0.04	-	-	-
		Model	0.48 \pm 0.18	1.02 \pm 0.71	0.000	***
CASE	Subject-independent	Baseline	0.17 \pm 0.05	-	-	-
		Model	0.22 \pm 0.05	0.44 \pm 0.94	0.001	***
	Subject-dependent (group)	Baseline	0.10 \pm 0.03	-	-	-
		Model	0.18 \pm 0.04	0.99 \pm 0.71	0.000	***
	Subject-dependent (subject)	Baseline	0.14 \pm 0.02	-	-	-
		Model	0.29 \pm 0.18	1.06 \pm 1.43	0.000	***
DREAMER	Subject-independent	Baseline	0.25 \pm 0.04	-	-	-
		Model	0.47 \pm 0.23	0.91 \pm 0.96	0.000	***
	Subject-dependent (group)	Baseline	0.20 \pm 0.05	-	-	-
		Model	0.39 \pm 0.07	1.09 \pm 0.61	0.000	***
	Subject-dependent (subject)	Baseline	0.24 \pm 0.05	-	-	-
		Model	0.48 \pm 0.18	1.08 \pm 0.82	0.000	***

Table E.40: Comparison of results in experimental setups (designs) with respective baselines (metric \pm std). Valence regression, average baseline (test). Note - gain represents scores relative to the baseline, averaged over a set of participants (may differ from simple calculation based on average values).

Dataset	Setup	Predictor	RMSE	Gain	p	Signif.
AMIGOS	Subject-independent	Baseline	0.26 \pm 0.07	-	-	-
		Model	0.43 \pm 0.26	0.75 \pm 1.20	0.000	***
	Subject-dependent (group)	Baseline	0.17 \pm 0.05	-	-	-
		Model	0.41 \pm 0.11	2.37 \pm 4.51	0.000	***
	Subject-dependent (subject)	Baseline	0.25 \pm 0.08	-	-	-
		Model	0.51 \pm 0.21	1.25 \pm 1.68	0.000	***
ASCERTAIN	Subject-independent	Baseline	0.28 \pm 0.05	-	-	-
		Model	0.42 \pm 0.13	0.50 \pm 0.44	0.000	***
	Subject-dependent (group)	Baseline	0.27 \pm 0.05	-	-	-
		Model	0.43 \pm 0.14	0.63 \pm 0.83	0.000	***
	Subject-dependent (subject)	Baseline	0.28 \pm 0.04	-	-	-
		Model	0.53 \pm 0.16	0.90 \pm 0.53	0.000	***
CASE	Subject-independent	Baseline	0.18 \pm 0.07	-	-	-
		Model	0.21 \pm 0.08	0.21 \pm 0.29	0.007	**
	Subject-dependent (group)	Baseline	0.11 \pm 0.04	-	-	-
		Model	0.18 \pm 0.06	0.77 \pm 0.39	0.000	***
	Subject-dependent (subject)	Baseline	0.15 \pm 0.03	-	-	-
		Model	0.36 \pm 0.22	1.50 \pm 1.55	0.000	***
DREAMER	Subject-independent	Baseline	0.32 \pm 0.03	-	-	-
		Model	0.41 \pm 0.08	0.29 \pm 0.26	0.000	***
	Subject-dependent (group)	Baseline	0.24 \pm 0.04	-	-	-
		Model	0.47 \pm 0.08	0.96 \pm 0.35	0.000	***
	Subject-dependent (subject)	Baseline	0.24 \pm 0.04	-	-	-
		Model	0.51 \pm 0.17	1.11 \pm 0.73	0.000	***

Table E.41: Results of Friedman test between experimental designs, based on relative differences between models and baselines.

Dataset	Task	χ^2	p (χ^2)	F	p (F)	Signif. (χ^2 / F)
AMIGOS classification	arousal	22.90	0.000	15.79	0.000	*** / ***
	valence	26.34	0.000	19.37	0.000	*** / ***
AMIGOS regression	arousal	4.15	0.125	2.14	0.124	ns / ns
	valence	19.54	0.000	12.70	0.000	*** / ***
ASCERTAIN classification	arousal	51.24	0.000	50.47	0.000	*** / ***
	valence	39.53	0.000	31.64	0.000	*** / ***
ASCERTAIN regression	arousal	20.04	0.000	12.22	0.000	*** / ***
	valence	25.92	0.000	17.04	0.000	*** / ***
CASE classification	arousal	18.20	0.000	12.63	0.000	*** / ***
	valence	12.60	0.001	7.71	0.001	** / **
CASE regression	arousal	21.67	0.000	16.39	0.000	*** / ***
	valence	31.67	0.000	32.41	0.000	*** / ***
DREAMER classification	arousal	21.48	0.000	19.27	0.000	*** / ***
	valence	18.09	0.000	14.26	0.000	*** / ***
DREAMER regression	arousal	3.13	0.209	1.61	0.212	ns / ns
	valence	18.87	0.000	15.30	0.000	*** / ***

Table E.42: Results of Conover's post-hoc test between experimental designs. AMIGOS dataset, Arousal classification. Comparisons were done on raw metrics. P-values were adjusted using Holm's procedure [209].

Group 1	Group 2	Statistic	p	Signif.
Subject-dependent (subject)	Subject-dependent (group)	-2.501	0.015	*
Subject-independent	Subject-dependent (group)	3.109	0.005	**
	Subject-dependent (subject)	5.609	0.000	***

Table E.43: Results of Conover's post-hoc test between experimental designs. AMI-GOS dataset, Valence classification. Comparisons were done on raw metrics. P-values were adjusted using Holm's procedure [209].

Group 1	Group 2	Statistic	p	Signif.
Subject-dependent (subject)	Subject-dependent (group)	-6.220	0.000	***
Subject-independent	Subject-dependent (group)	-2.903	0.005	**
	Subject-dependent (subject)	3.317	0.003	**

Table E.44: Results of Conover's post-hoc test between experimental designs. AMI-GOS dataset, Valence regression. Comparisons were done on raw metrics. P-values were adjusted using Holm's procedure [209].

Group 1	Group 2	Statistic	p	Signif.
Subject-dependent (subject)	Subject-dependent (group)	2.324	0.023	*
Subject-independent	Subject-dependent (group)	5.035	0.000	***
	Subject-dependent (subject)	2.711	0.017	*

Table E.45: Results of Conover's post-hoc test between experimental designs. AS-CERTAIN dataset, Arousal classification. Comparisons were done on raw metrics. P-values were adjusted using Holm's procedure [209].

Group 1	Group 2	Statistic	p	Signif.
Subject-dependent (subject)	Subject-dependent (group)	1.263	0.209	ns
Subject-independent	Subject-dependent (group)	9.264	0.000	***
	Subject-dependent (subject)	8.000	0.000	***

Table E.46: Results of Conover's post-hoc test between experimental designs. AS-CERTAIN dataset, Valence classification. Comparisons were done on raw metrics. P-values were adjusted using Holm's procedure [209].

Group 1	Group 2	Statistic	p	Signif.
Subject-dependent (subject)	Subject-dependent (group)	-7.516	0.000	***
Subject-independent	Subject-dependent (group)	-1.503	0.136	ns
	Subject-dependent (subject)	6.013	0.000	***

Table E.47: Results of Conover's post-hoc test between experimental designs. AS-CERTAIN dataset, Arousal regression. Comparisons were done on raw metrics. P-values were adjusted using Holm's procedure [209].

Group 1	Group 2	Statistic	p	Signif.
Subject-dependent (subject)	Subject-dependent (group)	-4.047	0.000	***
Subject-independent	Subject-dependent (group)	-4.484	0.000	***
	Subject-dependent (subject)	-0.437	0.663	ns

Table E.48: Results of Conover's post-hoc test between experimental designs. AS-CERTAIN dataset, Valence regression. Comparisons were done on raw metrics. P-values were adjusted using Holm's procedure [209].

Group 1	Group 2	Statistic	p	Signif.
Subject-dependent (subject)	Subject-dependent (group)	-4.314	0.000	***
Subject-independent	Subject-dependent (group)	1.249	0.215	ns
	Subject-dependent (subject)	5.562	0.000	***

Table E.49: Results of Conover's post-hoc test between experimental designs. CASE dataset, Arousal classification. Comparisons were done on raw metrics. P-values were adjusted using Holm's procedure [209].

Group 1	Group 2	Statistic	p	Signif.
Subject-dependent (subject)	Subject-dependent (group)	-5.018	0.000	***
Subject-independent	Subject-dependent (group)	-2.737	0.016	*
	Subject-dependent (subject)	2.281	0.026	*

Table E.50: Results of Conover's post-hoc test between experimental designs. CASE dataset, Valence classification. Comparisons were done on raw metrics. P-values were adjusted using Holm's procedure [209].

Group 1	Group 2	Statistic	p	Signif.
Subject-dependent (subject)	Subject-dependent (group)	-2.571	0.025	*
Subject-independent	Subject-dependent (group)	1.285	0.204	ns
	Subject-dependent (subject)	3.856	0.001	***

Table E.51: Results of Conover's post-hoc test between experimental designs. CASE dataset, Arousal regression. Comparisons were done on raw metrics. P-values were adjusted using Holm's procedure [209].

Group 1	Group 2	Statistic	p	Signif.
Subject-dependent (subject)	Subject-dependent (group)	1.588	0.118	ns
Subject-independent	Subject-dependent (group)	5.558	0.000	***
	Subject-dependent (subject)	3.970	0.000	***

Table E.52: Results of Conover's post-hoc test between experimental designs. CASE dataset, Valence regression. Comparisons were done on raw metrics. P-values were adjusted using Holm's procedure [209].

Group 1	Group 2	Statistic	p	Signif.
Subject-dependent (subject)	Subject-dependent (group)	-0.924	0.360	ns
Subject-independent	Subject-dependent (group)	6.465	0.000	***
	Subject-dependent (subject)	7.388	0.000	***

Table E.53: Results of Conover's post-hoc test between experimental designs. DREAMER dataset, Arousal classification. Comparisons were done on raw metrics. P-values were adjusted using Holm's procedure [209].

Group 1	Group 2	Statistic	p	Signif.
Subject-dependent (subject)	Subject-dependent (group)	-2.173	0.035	*
Subject-independent	Subject-dependent (group)	3.950	0.001	***
	Subject-dependent (subject)	6.123	0.000	***

Table E.54: Results of Conover's post-hoc test between experimental designs. DREAMER dataset, Valence classification. Comparisons were done on raw metrics. P-values were adjusted using Holm's procedure [209].

Group 1	Group 2	Statistic	p	Signif.
Subject-dependent (subject)	Subject-dependent (group)	-5.183	0.000	***
Subject-independent	Subject-dependent (group)	-1.481	0.146	ns
	Subject-dependent (subject)	3.702	0.001	**

Table E.55: Results of Conover's post-hoc test between experimental designs. DREAMER dataset, Valence regression. Comparisons were done on raw metrics. P-values were adjusted using Holm's procedure [209].

Group 1	Group 2	Statistic	p	Signif.
Subject-dependent (subject)	Subject-dependent (group)	-0.188	0.852	ns
Subject-independent	Subject-dependent (group)	4.694	0.000	***
	Subject-dependent (subject)	4.882	0.000	***