Investigating the Generalizability of Emotion Detection via Wearable Physiological Sensors: EmoWear Usecase

Aleksandr Ometov[*], Hsiao-Chun Lin[*], Anzhelika Mezina[†], Otso Arponen[‡],
Mohammad Hasan Rahmani[§¶], Kaarina Nikunen[*], Jari Nurmi[*]
[*]Tampere University, Tampere, Finland
[†]Brno University of Technology, Brno, Czech Republic
[‡]TAYS Hospital, Tampere, Finland
[§]CoSys-Lab, Faculty of Applied Engineering, University of Antwerp, Antwerp, Belgium
[¶]Flanders Make Strategic Research Centre, Lommel, Belgium
Email: aleksandr.ometov@tuni.fi

***Abstract*—Emotion detection is increasingly recognized as a foundational component in the evolution of intelligent eHealth systems. This paper presents a personalized methodology for identifying relevant physiological sensors for emotion recognition, using the open-access EmoWear dataset as a case study. The proposed framework applies statistical feature extraction and correlation analysis to examine the relationship between biosignals and emotional states, specifically valence, arousal, and dominance. The findings indicate that no single sensor modality universally correlates with emotional states across individuals, reinforcing the need for personalized and multimodal approaches. Notably, Electrodermal Activity (EDA) activity showed a higher correlation with valence, whereas Skin Temperature (SKT) was more closely associated with arousal; however, inter-individual variability remained significant. Overall, the analysis highlights challenges for generalizability in affective computing and emphasizes the importance of context and individual differences in emotion expression.**

***Index Terms*—Wearables, Emotion, Affection, Recognition, Biometry, Stress, Affective Computing**

I. INTRODUCTION

Ongoing advancement in wearable technologies and the proliferation of mobile sensing platforms have profoundly reshaped emotion recognition and mental health monitoring [1]. These technologies enable the collection of real-time physiological and behavioral data, providing the backbone for intelligent systems that detect emotional states with increasing accuracy and reliability.

Within the context of modern eHealth paradigms, emotion-aware systems play an increasingly important role in enhancing therapeutic interventions, especially for conditions where emotional well-being directly impacts clinical outcomes [2]. The rapid shift toward remote healthcare services, catalyzed by the COVID-19 pandemic, further reinforces the need to embed emotion recognition capabilities within telehealth platforms [3]. Such systems facilitate context-aware support, improve patient engagement, and offer timely insights into users' affective states.

Recent advancements in wearable devices have made it feasible to capture biosignals, e.g., Heart Rate Variability (HRV), Inter-Beat Interval (IBI) SKT-with high temporal resolution and minimal user burden, including in outdoor settings [4], [5] as well as indoor cases [6]. These data enable both self-monitoring and clinician-assisted evaluations. However, effectively leveraging these signals requires the development of robust Machine Learning (ML) pipelines capable of extracting reliable emotional insights under varying personal, and contextual differences [2].

Despite their promise, acquiring and using emotion-related data remain methodologically and ethically challenging [7]. Emotional states are highly subjective and vary across individuals, complicating the development of standardized data collection protocols [8]. Self-reported measures, while accessible, are susceptible to bias. By contrast, physiological indicators provide more objective inputs but demand careful calibration and interpretation [9], which developed a need for new ecosystems, see Fig. 1. Additionally, behavioral sensing methods, e.g., facial recognition and voice analysis raise substantial privacy and consent concerns under regulatory frameworks like the GDPR [3] and the EU AI Act [10].

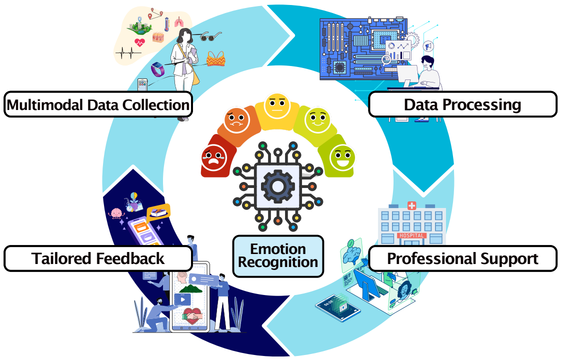


Fig. 1: Emotion detection and real-time feedback outlook

To address these limitations, multimodal approaches have gained attention and closer examinations. Integrating subjective, physiological, and behavioral modalities can improve both the robustness and generalizability of emotion recognition systems across diverse user groups and application scenarios [7]. Nevertheless, these approaches also increase complexity, raising novel questions about sensor relevance, data volume, and model interpretability.

This research aims to answer the following question: *Can physiological signals from wearable devices be used to infer emotional states across individuals with reasonable reliability?* To answer this question, we use the EmoWear dataset [11], which contains multimodal sensor data from two distinct wearable devices, for empirical validation. Its structure supports detailed analysis of individual sensor contributions to affective state inference. For a broader review of available multimodal datasets, refer to [12].

The remainder of this paper is structured as follows: Section II presents the dataset and extracted features. Section III outlines the analysis methodology. Section IV reports the key results. The last Section concludes the paper.

II. DATASET DESCRIPTION

This section provides a comprehensive description of the dataset and associated data modalities. For this study, the EmoWear dataset [11] was selected due to its multimodal nature and its collection using two distinct wearable devices: the Empatica E4 (wrist-based) and the Zephyr BioHarness 3 (chest-mounted). This dual-sensor configuration enables a more diverse and granular analysis of physiological and behavioral signals.

The dataset comprises recordings from 48 participants, incorporating the following sensor modalities (some collected simultaneously from two devices):

- **Empatica E4**: ACC, BVP, EDA, HR, IBI, and SKT.
- **Zephyr BioHarness 3 (BH3)**: ACC, BB, BR, ECG, HR, HR confidence, RR, and RSP.

Five affective state/emotion labels are included for each observation: valence (intrinsic attractiveness or averseness of an emotion), arousal (intensity or activation level of an emotional state), dominance (degree of control or influence one feels over the emotional situation), liking (degree of personal preference or fondness toward a stimulus or experience), and familiarity (how well known or recognized a stimulus or situation is to a person). These annotations follow a validated protocol for emotional self-assessment. All labels, with the exception of familiarity, are provided as continuous values in the range $[0,9]$, thus framing the primary task as a regression problem. Familiarity is a discrete variable, which may lend itself to classification-based treatment.

Each participant's recording is segmented into 38 time-synchronized sequences, with clearly defined label intervals. Using the provided timestamps, sequences are aligned to enable precise segmentation of continuous data into labeled windows. This structure benefits supervised learning by increasing the number of usable samples and facilitating sequence-aware modeling. Because sensors operate at different sampling frequencies, preprocessing is required (e.g., temporal alignment, interpolation, and resampling).

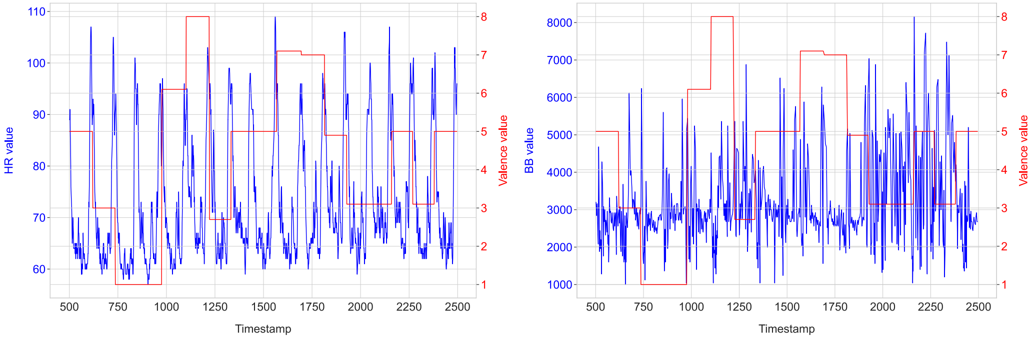
In the current analysis, the full dataset was initially examined to extract general insights. All statistical analyses were conducted on a per-participant basis to account for substantial inter-individual variability, motivating a personalized modeling approach.

A particular preprocessing note pertains to the Accelerometer (ACC) data, which consists of three axes: x, y, and z. A standard transformation involves computing the magnitude of the acceleration vector, defined as: $\vec{a} = \sqrt{x^2 + y^2 + z^2}$.

Further data processing steps and detailed analyses are outlined in the following subsections.

III. GENERAL ANALYSIS OF DATA

In this paper, an initial exploratory data analysis was conducted to investigate potential relationships between sensor signals and emotional labels. Figures 2a and 2b illustrate two such examples, overlaying HR and BB signals with the corresponding valence label for a representative participant. Visual inspection suggests that there are no immediately



(a) Heart Rate (HR) and Valence over Time

(b) Breath to Breath (BB) and Valence over Time

Fig. 2: Examples of Sensor Signals with Corresponding Valence Labels for participant 17-9UTL with normalised timestamps

discernible patterns, highlighting the complexity of emotion detection and the insufficiency of using a single signal to infer emotional states.

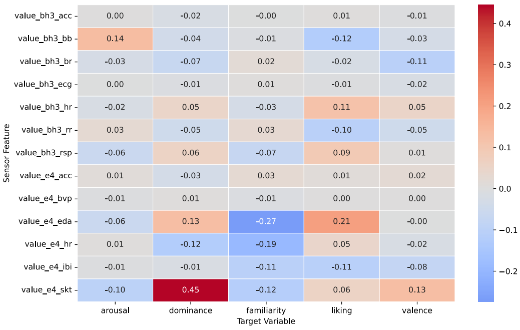
To statistically quantify the relationship between sensors and emotional labels, we employed the Spearman rank correlation coefficient [13], which is suitable for evaluating monotonic relationships between non-normally distributed variables. The coefficient is computed as [14]: $\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$, where d_i is the difference between the ranks of paired variables, and n is the number of samples.

Figures 3a and 3b present heatmaps of sensor-label correlations for two different participants. Notably, sensor-label correlations vary significantly across individuals. For example, participant 19-9UY7 exhibits meaningful correlation only between SKT and dominance, while participant 07-9UAD shows higher correlations between several sensors (e.g., EDA, SKT, Breath Rate (BR)) and multiple labels.

To generalize findings across participants, we examined boxplots of Spearman correlations. Figures 4, 5, and 6 present inter-sensor, sensor-label, and inter-label correlations, respectively. These visualizations assist in identifying consistent relationships and potential redundancies.

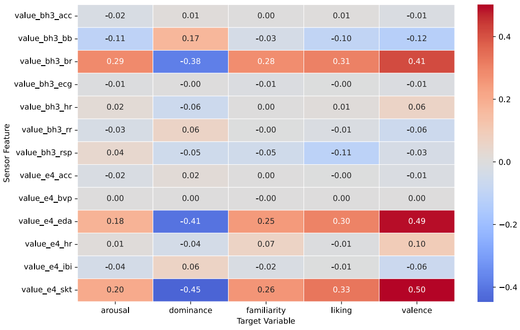
IV. MAIN FINDINGS ON THE DATA CORRELATION

a) Sensor-Label Relationships.: Fig. 4 reveals that certain sensors, e.g., *e4_acc*, *bh3_acc*, and *bh3_ecg*, exhibit negligible correlations with the emotional labels across par-



Sensor Feature	arousal	dominance	familiarity	liking	valence
value_bh3_acc	0.00	-0.02	-0.00	0.01	-0.01
value_bh3_bb	0.14	-0.04	-0.01	-0.12	-0.03
value_bh3_br	-0.03	-0.07	0.02	-0.02	-0.11
value_bh3_ecg	0.00	-0.01	0.01	-0.01	-0.02
value_bh3_hr	-0.02	0.05	-0.03	0.11	0.05
value_bh3_rr	0.03	-0.05	0.03	-0.10	-0.05
value_bh3_rsp	-0.06	0.06	-0.07	0.09	0.01
value_e4_acc	0.01	-0.03	0.03	0.01	0.02
value_e4_bvp	-0.01	0.01	-0.01	0.00	0.00
value_e4_eda	-0.06	0.13	-0.27	0.21	-0.00
value_e4_hr	0.01	-0.12	-0.19	0.05	-0.02
value_e4_ibi	-0.01	-0.01	-0.11	-0.11	-0.08
value_e4_skt	-0.10	0.45	-0.12	0.06	0.13

(a) Participant 19-9UY7: Sensor-Label Correlation



Sensor Feature	arousal	dominance	familiarity	liking	valence
value_bh3_acc	-0.02	0.01	0.00	0.01	-0.01
value_bh3_bb	-0.11	0.17	-0.03	-0.10	-0.12
value_bh3_br	0.29	-0.38	0.28	0.31	0.41
value_bh3_ecg	-0.01	-0.00	-0.01	-0.00	-0.01
value_bh3_hr	0.02	-0.06	0.00	0.01	0.06
value_bh3_rr	-0.03	0.06	-0.00	-0.01	-0.06
value_bh3_rsp	0.04	-0.05	-0.05	-0.11	-0.03
value_e4_acc	-0.02	0.02	0.00	-0.00	-0.01
value_e4_bvp	0.00	0.00	-0.00	0.00	0.00
value_e4_eda	0.18	-0.41	0.25	0.30	0.49
value_e4_hr	0.01	-0.04	0.07	-0.01	0.10
value_e4_ibi	-0.04	0.06	-0.02	-0.01	-0.06
value_e4_skt	0.20	-0.45	0.26	0.33	0.50

(b) Participant 07-9UAD: Sensor-Label Correlation

Fig. 3: Sensor-Label Correlation for Two Participants

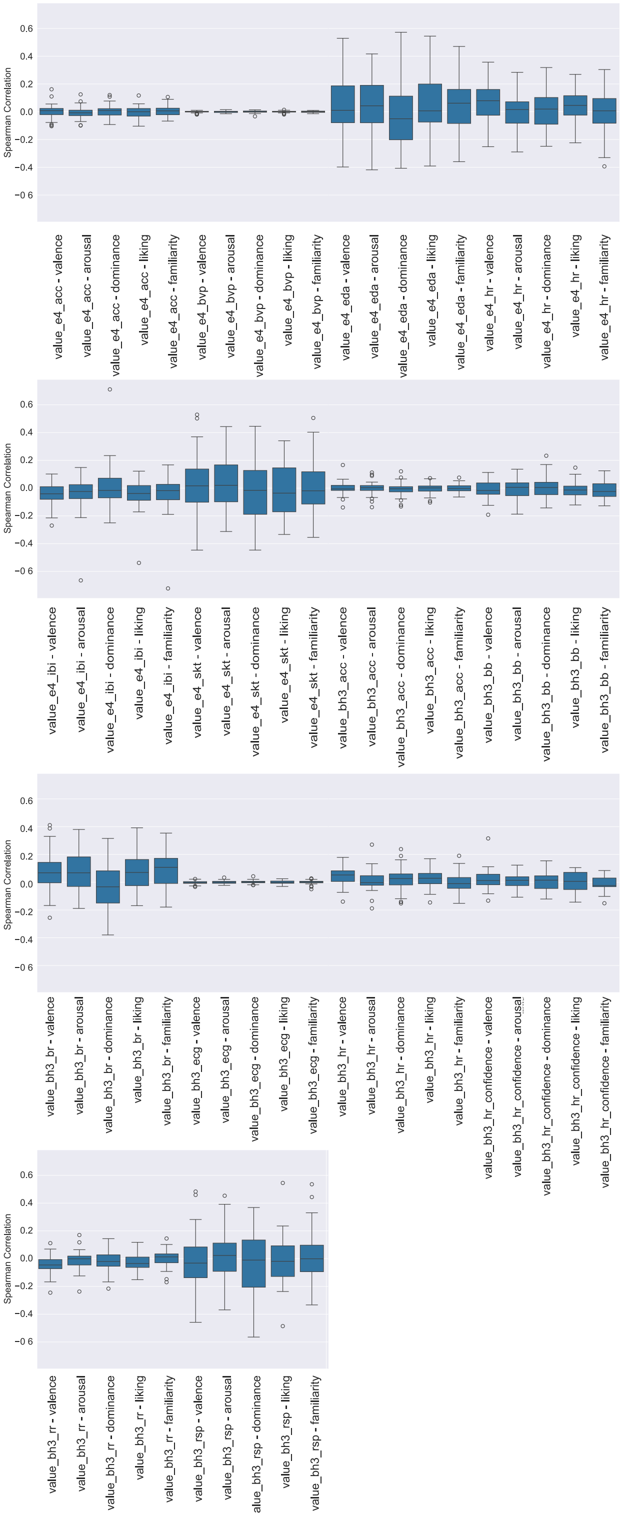


Fig. 4: Summarizing boxplots for pairs sensor-label

ticipants. This indicates that ACC and ElectroCardioGraphy (ECG) from the *bh3* may not reliably reflect the measured affective states. By contrast, *e4_eda* demonstrates considerably higher variability in its correlations with labels, consistent with its well-established physiological link to emotional arousal and stress responses. Additionally, breathing-related sensors display participant-dependent correlations, suggesting that the relationship between respiratory signals and affective states varies significantly across individuals. These findings highlight the need for personalization and adaptive modeling when using physiological data for emotion recognition, as uniform sensor-label relationships may not be true for all users.

b) Sensor-Sensor Relationships.: As shown in Fig. 5, correlations between sensor pairs vary, reflecting the diverse physiological processes captured by different modalities. For example, pairs *e4_bvp* and *e4_eda* exhibit consistently low correlations, indicating that they measure largely independent aspects of physiological responses. On the other hand, some sensor pairs are strongly correlated, particularly those derived from the same or related cardiac signals. Notably, *e4_ibi* and *bh3_hr*, as well as *bh3_hr* and *bh3_rr*, exhibit strong inverse correlations, which are physiologically expected due to their common cardiovascular origin and inverse relationship between IBI and HR. These results suggest that highly correlated signals might contain redundant information, and therefore, when applying ML methods, it may be beneficial to exclude such redundant features to reduce model complexity, improve interpretability, and enhance computational efficiency.

c) Label-Label Relationships.: As depicted in Fig. 6, the analysis reveals meaningful relationships among affective labels. The strongest positive correlation is observed between valence and liking, suggesting that participants generally tend to favor stimuli they perceive as pleasant or positive, reinforcing the intuitive link between emotional positivity and preference. Similarly, liking shows a positive correlation with familiarity, indicating that more familiar stimuli are generally liked more. However, empirical associations can arise from stimulus selection and reporting. In DEAP dataset [15], videos were chosen to maximize spread in the valence–arousal space, with no comparable constraint on dominance; this design choice may have effects on observed relationships. In our data, arousal and dominance demonstrate an inverse relationship, where higher arousal tends to coincide with lower feelings of control or dominance, which aligns with psychological theories of emotional experience. Other label pairs exhibit weak correlation, underscoring the complexity and multidimensional nature of affective states and emphasizing the importance of considering multiple emotional dimensions in comprehensive emotion recognition systems. Taken together, these observed correlations likely reflect both genuine psychological regularities and dataset-induced sampling effects and should therefore be interpreted with appropriate caution.

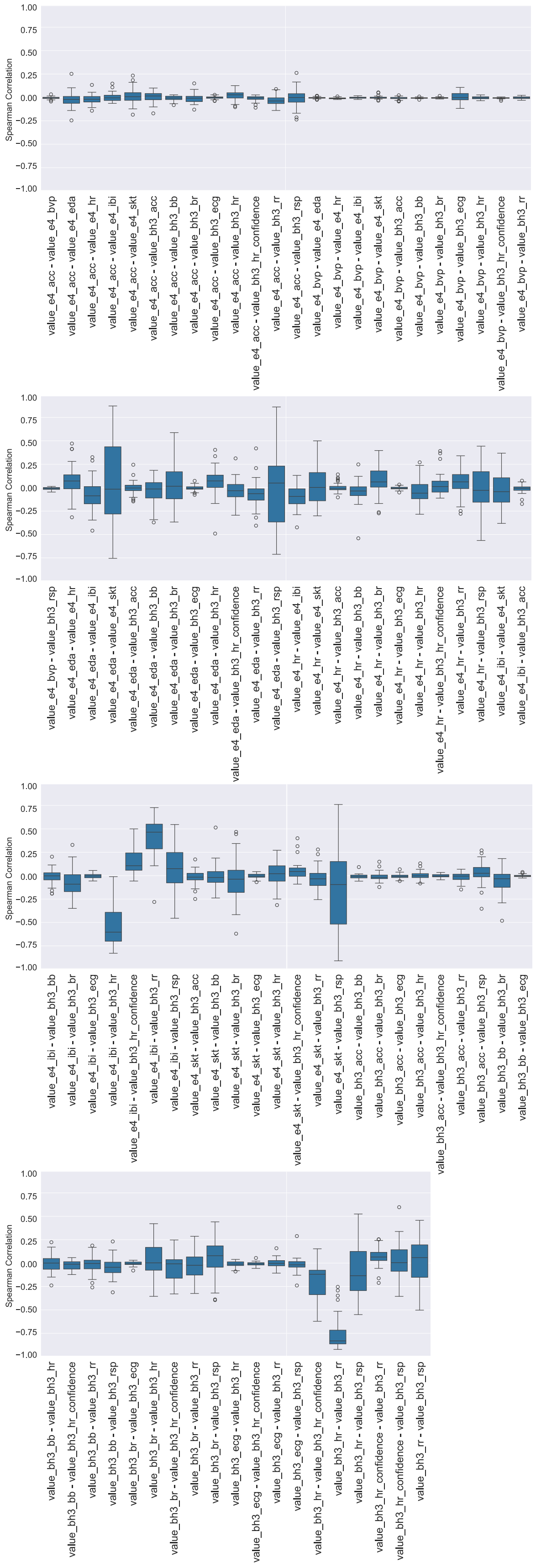


Fig. 5: Summarizing boxplots for pairs sensor-sensor

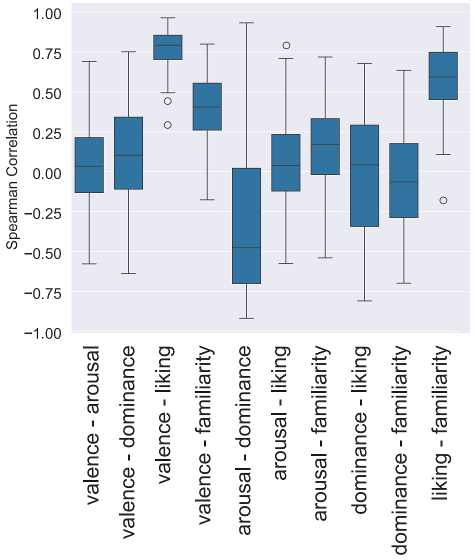


Fig. 6: Summarizing boxplots for pairs label-label

d) Key Observations: From this comprehensive analysis and Fig. 7, several important conclusions can be drawn that have implications for future research and practical applications. First, no single sensor consistently provides sufficient predictive power across the entire participant cohort, indicating that relying on just one physiological or behavioral signal may not capture the full complexity of emotional states. Second, some sensor-label relationships, most notably the association between EDA and arousal, are both physiologically plausible and statistically robust, confirming the validity of using certain biosignals as reliable indicators of specific affective dimensions. Lastly, the pronounced inter-individual variability observed across multiple sensor signals and their correlations with emotional labels strongly underscores the necessity of adopting personalized modeling strategies. Such approaches can better accommodate unique physiological and psychological patterns, thereby improving the accuracy and relevance of emotion recognition systems for individual users.

These findings guide the subsequent feature engineering and modeling efforts, suggesting that both general trends and individual signal-profiles must be considered in the development of effective emotion recognition systems.

V. DISCUSSION AND CONCLUSION

This study explored emotion recognition from physiological signals gathered from wearable devices, using the EmoWear dataset. Through a combination of statistical analysis and exploratory data visualization, we investigated the complex relationship between sensor readings and emotional labels, e.g., valence, arousal, dominance, liking, and familiarity. The analysis revealed that no single sensor consistently correlates with emotional states across all participants. While certain modalities, i.e., EDA, SKT, and BR occasionally exhibited meaningful associations with specific labels, these correlations varied significantly between individuals.

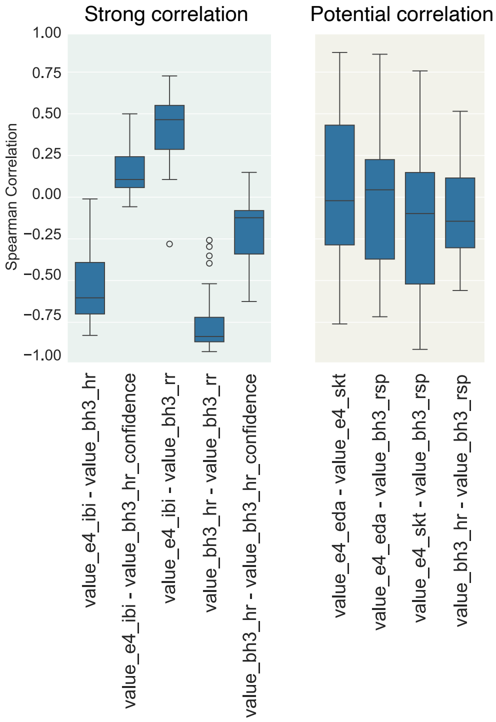


Fig. 7: Sensor-sensor and label-sensor pairs with strong or potential correlation

These findings highlight the deeply personal nature of physiological responses to emotional stimuli, reinforcing the necessity of individualized modeling rather than generalized approaches. The observed variability in correlation patterns across participants indicates that affective states cannot be reliably inferred from isolated sensor data; instead, a multimodal fusion of signals is required. Sensor-sensor analysis revealed both expected and unexpected dependencies, which can inform more efficient feature selection by identifying redundant or non-informative inputs.

The data-driven insights also have implications for the development of emotion-aware systems. Although statistical methods can guide the reduction of dimensionality and inform feature engineering, the broader context of human variability and the ethical dimensions of emotion detection must also be considered. Emotions are complex, socially mediated phenomena, and any technological solutions intended to recognize them should respect their subjectivity and address privacy concerns inherent to physiological monitoring.

Overall, the findings support a modeling strategy that combines multimodal sensor input with a personalized analysis framework. We emphasize, however, that these findings are restricted by the sensor, examination protocol, and sample characteristics of EmoWear. The broader claims are required for replication, so that such an approach is likely to yield more accurate and contextually relevant emotion recognition systems. Future work will build on this foundation, integrating temporal models and exploring real-time feedback mechanisms, while continuing to evaluate the ethical implications of deploying these technologies in real-world settings.

ACKNOWLEDGMENT

H.-C. Lin, A. Mezina contributed equally to this research. This research was supported by Jane and Aatos Erkko Foundation through CONVERGENCE of Humans and Machines project. Copilot was used for additional proofreading and grammar checks.

REFERENCES

[1] D. Nandini, J. Yadav, V. Singh, V. Mohan, and S. Agarwal, “An Ensemble Deep Learning Framework for Emotion Recognition through Wearable Devices Multi-Modal Physiological Signals,” *Scientific Reports*, vol. 15, no. 1, p. 17263, 2025.

[2] R. Guo, H. Guo, L. Wang, M. Chen, D. Yang, and B. Li, “Development and Application of Emotion Recognition Technology – A Systematic Literature Review,” *BMC Psychology*, vol. 12, no. 95, pp. 1–15, 2024.

[3] R. Pereira, C. Mendes, J. Ribeiro, R. Ribeiro, R. Miragaia, N. Rodrigues, N. Costa, and A. Pereira, “Systematic Review of Emotion Detection with Computer Vision and Deep Learning,” *Sensors*, vol. 24, no. 11, p. 3484, 2024.

[4] T. Kim, S. Kim, M. Lee, Y. Kang, and S. Hwang, “Assessing Human Emotional Experience in Pedestrian Environments Using Wearable Sensing and Machine Learning with Anomaly Detection,” *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 109, pp. 540–555, 2025.

[5] H.-C. Lin, A. Ometov, O. Arponen, K. Nikunen, and J. Nurmi, “Stress DeTech-tion: Revolutionizing Wellbeing in Future Networks,” in *Proceedings of the 2024 International Conference on Information Technology for Social Good*, 2024, pp. 114–117.

[6] A. Site, E. S. Lohan, O. Jolanki, O. Valkama, R. R. Hernandez, R. Latikka, D. Alekseeva, S. Vasudevan, S. Afolaranmi, A. Ometov *et al.*, “Managing Perceived Loneliness and Social-Isolation Levels for Older Adults: A Survey with Focus on Wearables-based Solutions,” *Sensors*, vol. 22, no. 3, p. 1108, 2022.

[7] H.-C. Lin, A. Ometov, O. Arponen, K. Nikunen, and J. Nurmi, “Stress Monitoring in the Era of AI Boom – Where Should We Begin?” in *Proceedings of the 2025 ACM 5th International Conference on Information Technology for Social Good*, 2025.

[8] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression,” *Proc. of IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 94–101, 2010.

[9] S. D’Mello and J. Kory, “A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion,” *IEEE Tr. on Aff. Comp.*, vol. 7, no. 1, pp. 18–31, 2015.

[10] European Commission, “Annex to the Communication to the Commission – Approval of the content of the draft Communication from the Commission – Commission guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act), C(2025) 884 final,” 2025, accessed: 2025-06-10.

[11] M. H. Rahmani, M. Symons, O. Sobhani, R. Berkvens, and M. Weyn, “EmoWear: Wearable Physiological and Motion Dataset for Emotion Recognition and Context Awareness,” *Scientific Data*, vol. 11, no. 1, p. 648, 2024.

[12] A. Ometov, A. Mezina, H.-C. Lin, O. Arponen, R. Burget, and J. Nurmi, “Stress and Emotion Open Access Data: A Review on Datasets, Modalities, Methods, Challenges, and Future Research Perspectives,” *J. of Healthcare Informatics Research*, 2025.

[13] M. M. Mukaka, “A Guide to Appropriate Use of Correlation Coefficient in Medical Research,” *Malawi Medical J.*, vol. 24, no. 3, pp. 69–71, 2012.

[14] H. Yu and A. D. Hutson, “A Robust Spearman Correlation Coefficient Permutation Test,” *Communications in Statistics-Theory and Methods*, vol. 53, no. 6, pp. 2141–2153, 2024.

[15] 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 Tr. on Aff. Comp.*, vol. 3, no. 1, pp. 18–31, 2011.