Participant-Independent Prediction of Self-Reported Valence and Arousal in Virtual Reality Empathy Scenes Using EmteqPRO Sensor Data

Viktorija Projkova, Kristina Kitrozoska, Emilija Kizevska
Martina Miskova, Teodora Domazetovic, Hristijan Gjoreski

Faculty of Electrical Engineering and Information Technologies
University of Ss. Cyril and Methodius in Skopje, Macedonia
projkova.viktorija@gmail.com, martinamiskova15@gmail.com, kristinakitroz@gmail.com,
tdomazetovic13@gmail.com, ekizevska@feit.ukim.edu.mk, hristijang@feit.ukim.edu.mk

Abstract

Valence reflects the pleasantness-unpleasantness dimension of affect, while arousal represents activation intensity. Using four virtual reality (VR) empathy scenes, this study evaluates whether EmteqPRO physiological, motion, and facial features can predict self-reported valence and arousal in unseen participants. Of 105 volunteers, 96 passed quality control. Data were collected with a Pico Neo 3 Pro Eye headset using facial EMG, PPG, and IMU sensors, together with device-produced facial activation and expression-intensity features. Friedman and Wilcoxon tests assessed scene differences, while two-layer feature-group stacking classified lower versus higher ratings under participant-independent Leave-One-Participant-Out (LOPO) validation. The best self-reported valence configuration achieved 0.705 balanced accuracy and 0.683 macro F1-score. The best feature-group stacking configuration for self-reported arousal achieved 0.462 balanced accuracy and macro F1-score, while the direct all-feature Extra Trees baseline achieved 0.533 on both metrics.

Keywords

virtual reality, wearable sensors, affective computing, valence, arousal, multimodal machine learning

1 Introduction

Quantification of human affect is crucial for Human-Computer Interaction (HCI) and affect-aware environments [1]. Although machine learning can monitor affect-related properties through physiological signals, models that generalize to new users remain challenging because responses vary across people and contexts [2]. Recordings were collected from 105 participants using the EmteqPRO device and Pico Neo 3 Pro Eye headset following a previously published VR empathy protocol [3], of whom 96 were retained after quality control in the released dataset [4]. Facial EMG, PPG, and IMU signals, together with device-produced facial outputs, form five model input feature groups. One first-layer model is trained per group and combined through stacked generalization [5].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.skui.74

The main contribution is feature-group stacking for participant-independent prediction of self-reported valence and arousal, with a leakage-controlled comparison of classifiers and sampling strategies. Performance was evaluated using LOPO cross-validation to assess generalization to unseen participants.

Earlier affect-recognition studies used EMG and autonomic responses [1,6], while multimodal work combined facial-muscle, cardiac, and respiratory features [7]. Prior studies did not jointly evaluate feature-group stacking for participant-independent prediction of self-reported valence and arousal in immersive VR.

Physiological responses are context- and participant-dependent, complicating unseen-user generalization. Recent VR empathy studies using EmteqPRO sensor data have highlighted the importance of multimodal modelling and participant-independent evaluation when estimating empathy and affective states [8,9].

2 Data

The dataset contained EmteqPRO recordings and scene-level questionnaires from 105 VR participants using the Pico Neo 3 Pro Eye headset [4]. Quality control retained 96 participants and 384 samples, with four scenes per participant. Seven participants lacked reliable video timing and two had an empathy scene without usable core model input features. Data collection followed an established VR empathy protocol combining VR stimuli, sensor recordings, and self-reported mental-state ratings [8].

A session is one participant's complete VR experiment. Within a session, the continuous VR video sequence comprised six scene-specific segments: a forest baseline, four empathy scenes, and a roller coaster. A scene is one VR stimulus, while a segment is the corresponding temporal portion of the continuous sequence. Only the four empathy scenes were analyzed; the forest served as a resting baseline and the roller coaster as a non-empathic arousal control. Each empathy scene showed an actor portraying anger, sadness, happiness, or anxiousness according to the experimental script [3]. These labels refer to the emotion portrayed by the actor and do not imply that the participant experienced the same categorical emotion. No independent categorical-emotion validation was collected; the labels were therefore used as scene identifiers.

The physical sensors were facial EMG, PPG, and IMU [11]. EmteqPRO processing derived facial-muscle amplitudes from

EMG, HRV and breathing measures from PPG, and motion intensity from IMU; the Emotion AI Engine generated facial activation and expression-intensity outputs [4]. Device-produced arousal and valence were Emotion AI Engine estimates derived from four-feature EMG/PPG fusion, while facial valence was an EMG-based facial-affect estimate [4]. These device-produced affect outputs were excluded from model inputs to avoid leakage with the self-reported targets.

The prediction targets were ordinal self-reported valence and arousal collected after the four empathy scenes. Valence refers to pleasantness-unpleasantness, while arousal represents activation intensity [10]. Discomfort was exploratory only. Each sample, one participant–empathy-scene pair, contained metadata, targets, and 36 minimum or maximum model input features.

3 Methodology

3.1 Data Preprocessing and Target Definition

EmteqPRO timestamps were aligned with questionnaire and timing metadata. The video was divided into six segments, and 40 s were trimmed from each boundary to reduce transition and questionnaire noise, with shorter trimming where needed to retain at least 30 s. The resulting interval defined the analysis window, and only the four empathy-scene windows were used. Scenes with unreliable timing or missing core inputs were excluded. Window minima and maxima were extracted, absent expression intensities were set to zero, and quality-control and direct target outputs were excluded.

The 36 minimum/maximum columns were derived from 18 original variables ($18 \times 2 = 36$). For each variable, midrange = (minimum + maximum)/2 and range = maximum − minimum were added, yielding 72 target-independent features ($18 \times 4 = 72$). Because midrange and range are deterministic functions of the extrema, they introduce feature redundancy. Identifiers, metadata, targets, timing information, and quality-control fields were excluded from the model feature matrix.

The original self-reported valence and arousal ratings used a five-point scale from 0 to 4. For valence, lower values represented greater pleasantness and higher values greater unpleasantness; for arousal, lower values represented lower activation and higher values greater activation. Ratings 0–1 formed the lower-score range, rating 2 the intermediate category, and ratings 3–4 the higher-score range. The central rating was kept separate, and for binary classification it was excluded rather than reassigned to either class, yielding a clearer comparison between lower and higher numerical ratings.

3.2 Statistical Methodology

Complete participant-level four-scene vectors were analyzed. Friedman tests [12] assessed scene effects in the targets and selected inputs. Each significant Friedman result underwent six Wilcoxon comparisons [13]. Bonferroni-adjusted p-values were reported and significance was assessed at adjusted $p<0.05$ [14]. Before testing, one representative feature was selected from each signal group for the statistical scene comparisons only, separate from the full 72-feature set used as machine-learning input, to provide an interpretable overview while limiting the number of comparisons. Minimum breathing rate represented respiratory activity, maximum RR interval (the time between successive heartbeats) cardiac/HRV-related timing, maximum zygomaticus EMG facial-muscle activation, maximum smile intensity expression-related behaviour, and maximum facial activation overall device-produced facial activity. Tests used original values, while standardized means were used only for descriptive visualization.

3.3 Machine Learning Methodology

The 72 features from Section 3.1 were partitioned into five first-layer feature groups of related signals or outputs. Five group-specific models formed the first layer; a Logistic Regression meta-model combined their five predicted probabilities for the higher-score class, P(high), in the second layer. Facial activation contained minimum, maximum, midrange, and range (4 features). EMG activation contained four summaries for zygomaticus, orbicularis, frontalis, and corrugator (16); expressions for overall, neutral, smile, eyebrow raise, and frown (20); breathing for breathing rate and stream-specific IMU motion intensity (8); HRV for mean heart rate, RR interval, standard deviation of normal-to-normal intervals (SDNN), standard deviation of successive differences (SDSD), root mean square of successive differences (RMSSD), and a separate stream-specific IMU motion intensity (24). The two IMU variables came from different streams, so no feature was duplicated across models.

In every inner and outer split, first-layer pipelines applied median imputation, VarianceThreshold, and StandardScaler [15], fitted only on training data. Logistic Regression, Extra Trees [16], and XGBoost [17] each returned P(high). Random Oversampling (ROS) [18] was included to address class imbalance by increasing the representation of minority-class samples in the training data. To keep the experimental comparison focused, ROS was evaluated only with the two tree-based classifiers, Extra Trees and XGBoost, while Logistic Regression was retained as a class-weighted linear baseline. ROS was applied separately within each feature-group pipeline and only to training samples. The class-balanced Logistic Regression meta-model and its scaler were fitted only on meta-training data.

Hyperparameters were fixed across folds. Logistic Regression used balanced class weights and a maximum of 5000 iterations. Extra Trees used 300 estimators, a minimum leaf size of 2, and balanced class weights. XGBoost used 200 estimators, a learning rate of 0.05, maximum depth of 2, row and column subsampling of 0.8, and log-loss evaluation. All stochastic components used a random seed of 42 where applicable, and ROS used the default sampling strategy.

Nested participant grouping generated leakage-free meta-features, consistent with participant-independent evaluation in related VR empathy modelling work [9]. In each outer LOPO fold, one participant was held out. An inner five-fold GroupKFold [15] on the remaining participants produced out-of-

fold P(high) values for the five-column meta-training matrix. Base models were refitted on all outer-training participants to generate held-out inputs, after which the meta-model predicted the unseen participant. Preprocessing, ROS, base-model fitting, and meta-model training used only training participants in each outer fold. Outer LOPO was repeated separately for self-reported valence and arousal using participant ID. Accuracy, balanced accuracy, macro F1-score, and weighted F1 were reported. Balanced accuracy and macro F1-score were the primary comparison metrics.

4 Experimental Setup

4.1 Statistical Analysis

Statistical experiments evaluated scene effects in self-reported valence and arousal and in minimum breathing rate, maximum zygomaticus EMG amplitude, RR interval, smile intensity, and facial activation across the four empathy scenes. The analysis followed the tests and significance criteria in Section 3.2.

4.2 Machine Learning Experiments

Five feature-group stacking configurations were evaluated separately for binary self-reported valence and arousal: Logistic Regression without sampling, and Extra Trees and XGBoost with and without ROS. All used the five feature groups, a Logistic Regression meta-classifier, and participant-independent evaluation from Section 3.3. Baselines included direct Extra Trees on all 72 features, unweighted probability averaging, dummy and scene-only classifiers; valence stacking was also summarized by scene.

5 Results

5.1 Statistical Analysis Results

Figures 1 and 2 show visible scene-dependent differences in self-reported valence and arousal, with the happiness scene showing the lowest ratings for both affective dimensions.

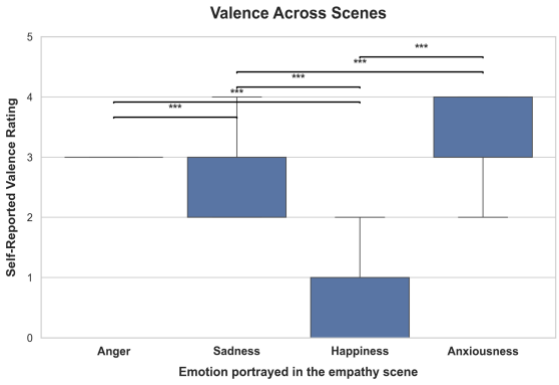


Figure 1: Self-reported valence ratings across the four empathy scenes

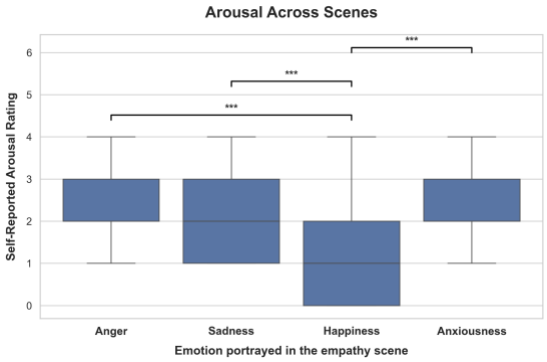


Figure 2: Self-reported arousal ratings across the four empathy scenes

Friedman tests showed significant scene effects for self-reported valence, $\chi^2(3)=201.165$, $p<0.001$, and arousal, $\chi^2(3)=83.388$, $p<0.001$ (Figures 1–2). Bonferroni-adjusted Wilcoxon tests found five significant valence contrasts ($W=92.0–319.0$, all $p<0.001$), with no difference between anger and anxiousness ($W=308.0$, $p=1.000$). For arousal, significant contrasts involved happiness versus anger, sadness, and anxiousness ($W=266.0–369.0$, all $p<0.001$), with happiness lowest. Significant scene effects were also observed for minimum breathing rate, $\chi^2(3)=9.417$, $p=0.024$, maximum zygomaticus EMG, $\chi^2(3)=67.025$, $p<0.001$, smile intensity, $\chi^2(3)=70.323$, $p<0.001$, and facial activation, $\chi^2(3)=52.738$, $p<0.001$, but not maximum RR, $\chi^2(3)=2.740$, $p=0.433$. Post-hoc tests showed an anger–anxiousness difference in minimum breathing rate ($W=1339.0$, adjusted $p=0.0029$), while maximum zygomaticus EMG, smile intensity, and facial activation differed for anger–happiness, anger–anxiousness, sadness–happiness, and sadness–anxiousness ($W=161.0–1207.0$, all adjusted $p<0.001$).

5.2 Machine Learning Results

After excluding intermediate ratings, self-reported valence contained 302 samples (88 lower-score, 214 higher-score) and self-reported arousal 253 (121 lower-score, 132 higher-score).

Table 1: Balanced accuracy (BA) and macro F1-score for self-reported valence and arousal by feature-group stacking configuration

Configuration	Valence BA	Valence F1	Arousal BA	Arousal F1
Logistic Regression	0.679	0.658	0.462	0.462
Extra Trees	0.705	0.683	0.426	0.426
Extra Trees + ROS	0.666	0.645	0.421	0.421
XGBoost	0.675	0.658	0.426	0.426
XGBoost + ROS	0.677	0.648	0.429	0.429

Table 1 summarizes stacking performance. Extra Trees without ROS was best for valence (BA 0.705, macro F1 0.683); Logistic Regression without ROS was best among arousal stacking configurations (0.462 for both metrics). Baselines yielded BA/macro F1 of 0.650/0.657 (direct Extra Trees) and 0.601/0.607 (unweighted averaging) for valence, and 0.533/0.533 and 0.492/0.492 for arousal; dummy BA was 0.500 for both and scene-only BA 0.500/0.553. Figure 3 shows the best stacking-model confusion matrices: valence correctly classified 60/88 lower-score and 156/214 higher-score samples (recall 0.68/0.73), and arousal 54/121 and 63/132 (0.45/0.48).

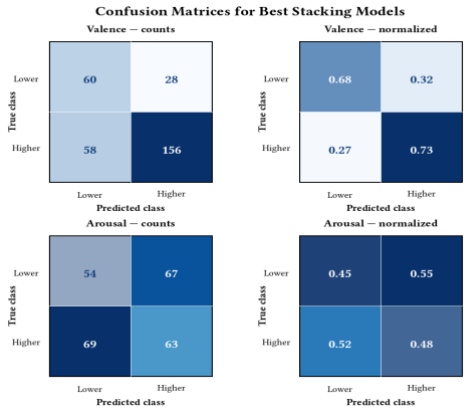


Figure 3: Confusion matrices for the best-performing feature-group stacking models for self-reported valence and arousal

Predicted probabilities showed clearer class separation for self-reported valence than for self-reported arousal, whose probabilities overlapped near 0.5 (Figure 4).

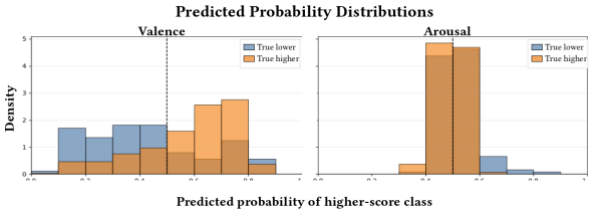


Figure 4: Predicted probability distributions for the best-performing self-reported valence and arousal feature-group stacking models

5.3 Discussion

The participant-independent models predicted self-reported valence more reliably than arousal, consistent with the weaker scene separation observed for arousal. This may reflect overlapping and individually variable physiological responses [2,10]. Arousal probabilities clustered near 0.5, while facial activation and expression-intensity features may have supported more consistent valence patterns. Valence stacking varied across scenes (BA 0.630–0.735), despite chance-level valence scene-only performance, while the arousal scene-only baseline reached 0.553 BA, suggesting that scene identity carried some predictive information. ROS did not consistently improve performance, indicating that duplicating minority-class samples provided little benefit for generalization to unseen participants. Overall, stacking generalized reasonably for valence, whereas arousal remained difficult and may require participant-aware modelling.

6 Conclusion

This study evaluated participant-independent prediction of self-reported valence and arousal from multimodal EmteqPRO features collected during VR empathy scenes. Statistical analysis confirmed scene-dependent differences, but these alone did not establish predictive performance across unseen participants.

Feature-group stacking generalized reasonably for valence, with Extra Trees reaching 0.705 balanced accuracy and 0.683 macro F1-score. Arousal remained more difficult: the best stacking configuration achieved 0.462 on both metrics, while the direct all-feature Extra Trees baseline reached 0.533. These results show that the same multimodal representation did not generalize equally across affective dimensions. Future work should explore personalized or participant-aware modelling, particularly for arousal.

References

[1] J. W. Tan, A. O. Andrade, H. Li, S. Walter, D. Hrabal, S. Rukavina, et al. 2016. Recognition of intensive valence and arousal affective states via facial electromyographic activity in young and senior adults. PLOS ONE 11, 1 (2016), e0146691. DOI: https://doi.org/10.1371/journal.pone.0146691.

[2] M. Gendron, C. Crivelli, and L. F. Barrett. 2018. Universality reconsidered: Diversity in making meaning of facial expressions. Current Directions in Psychological Science 27, 4 (2018), 211–219. DOI: https://doi.org/10.1177/0963721417746794.

[3] E. Kizhevska, K. Špareblek, and M. Luštrek. 2024. Protocol of the study for predicting empathy during VR sessions using sensor data and machine learning. PLOS ONE 19, 7 (2024), e0307385. DOI: https://doi.org/10.1371/journal.pone.0307385.

[4] E. Kizhevska, K. Špareblek, and M. Luštrek. 2025. Sensor data over 100 participants recorded with EmteqPRO and Pico 3 VR headset during empathy elicitation. IEEE DataPort. [Online]. Available: https://ieee-dataport.org/documents/sensor-data-over-100-participants-recorded-emteqpro-pico-3-vr-headset-during-empathy [Accessed: Jun. 21, 2026].

[5] D. H. Wolpert. 1992. Stacked generalization. Neural Networks 5, 2 (1992), 241–259. DOI: https://doi.org/10.1016/S0893-6080(05)80023-1.

[6] K. Wolf. 2015. Measuring facial expression of emotion. Dialogues in Clinical Neuroscience 17, 4 (2015), 457–462. DOI: https://doi.org/10.31887/DCNS.2015.17.4/kwolf.

[7] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. 2018. Multimodal assessment of depression from behavioral signals. In Handbook of Multimodal-Multisensor Interfaces, Volume 2. ACM and Morgan & Claypool, 375–417.

[8] E. Kizhevska, H. Gjoreski, and M. Luštrek. 2025. Predicting empathy and other mental states during VR sessions using sensor data and machine learning. Sensors 25, 18 (2025), 5766. DOI: https://doi.org/10.3390/s25185766.

[9] E. Kizhevska, H. Gjoreski, and M. Luštrek. 2026. A Two-Stage Methodology Combining Clustering and Predictive Models to Estimate the Levels of Participants' Empathy and Other Affective States During Virtual Reality Sessions. IEEE Access 14 (2026), 38493–38502. DOI: https://doi.org/10.1109/ACCESS.2026.3668967.

[10] J. A. Russell and L. F. Barrett. 1999. Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. Journal of Personality and Social Psychology 76, 5 (1999), 805–819. DOI: https://doi.org/10.1037/0022-3514.76.5.805.

[11] Emteq Labs. 2025. Pico Neo 3 Pro Eye Guide. Emteq Labs Support Docs. [Online]. Available: https://support.emteqlabs.com/02_emteqpro/01_quick-start-guides/quick-start-guide-pico-neo3.html [Accessed: Jun. 21, 2026].

[12] M. Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32, 200 (1937), 675–701. DOI: https://doi.org/10.1080/01621459.1937.10503522.

[13] F. Wilcoxon. 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1, 6 (1945), 80–83. DOI: https://doi.org/10.2307/3001968.

[14] O. J. Dunn. 1961. Multiple comparisons among means. Journal of the American Statistical Association 56, 293 (1961), 52–64. DOI: https://doi.org/10.1080/01621459.1961.10482090.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12 (2011), 2825–2830.

[16] P. Geurts, D. Ernst, and L. Wehenkel. 2006. Extremely randomized trees. Machine Learning 63, 1 (2006), 3–42. DOI: https://doi.org/10.1007/s10994-006-6226-1.

[17] T. Chen and C. Guestrin. 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 785–794. DOI: https://doi.org/10.1145/2939672.2939785.

[18] G. Lemaître, F. Nogueira, and C. K. Aridas. 2017. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research 18, 17 (2017), 1–5.