Predicting Mental States During VR Sessions Using Sensor Data and Machine Learning

Emilija Kizhevska*
emilija.kizhevska@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School (IPS)
Ljubljana, Slovenia

Mitja Luštrek
mitja.lustrek@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School (IPS)
Ljubljana, Slovenia

Abstract

Empathy is a multifaceted concept with both cognitive and emotional components that plays a crucial role in social interactions, prosocial behavior, and mental health. In our study, empathy and general arousal were induced via VR, with physiological signals measured and ground truth collected through questionnaires. Data from over 100 participants were collected and analyzed using multiple machine learning models and classification algorithms to predict empathy based on physiological responses. We explored different data balancing techniques and labeled data in multiple ways to enhance model performance. Our results show that they are effective in detecting general arousal, empathy, and differentiating between non-empathic and empathic arousal, but the models encountered difficulties with precise emotion detection. The dataset extracted at 5-second intervals and models using Random Forest and Extreme Gradient Boosting showed the best performance. Future work will focus on refining emotion detection through advanced modeling techniques and investigating gender differences in empathy.

Keywords

VR, mental states, machine learning, sensor data

1 Introduction

Empathy is a multifaceted concept explored across various fields, including psychology, neuroscience, and sociology. Though no universal definition exists, empathy is generally understood to include both cognitive (understanding another's perspective) and emotional (experiencing another's feelings) components [8]. Our research defines empathy as the ability to model others' emotional states and respond sensitively while recognizing the self-other distinction [14].

There is no "golden standard" for measuring empathy [10], with methods varying from self-report questionnaires to psychophysiological measures like heart rate and skin conductance. Each method has its pros and cons, often leading to a combination of approaches for a comprehensive assessment. Psychophysiological measures offer objective data but face challenges due to individual variability and non-empathetic factors. Our study addresses these issues by using machine learning to directly measure empathy from physiological signals, offering a novel approach.

VR creates an immersive environment that enhances empathy by allowing users to experience different perspectives and engage emotionally. VR is effective for empathy training and is referred to as 'the ultimate empathy machine' [1, 11] for various reasons: 1) Immersive Experience: Provides a strong sense of presence, helping users adopt new viewpoints [15]. 2) Perspective-Taking and Emotional Engagement: Simulates realistic scenarios to provoke emotional responses and understanding [19]. 3) Empathy Training: Effective in healthcare, education, and diversity training by challenging preconceptions and deepening emotional insights [16]. 4) Ethical Considerations: Ensures respectful use of VR, balancing immersive experiences with participants' well-being [2].

The objective of this study was to examine how participants' empathy correlates with changes in their physiological metrics, measured using sensors such as inertial measurement unit (IMU), photoplethysmograph (PPG), and electromyography (EMG). Participants were immersed in 360° VR videos featuring actors displaying various emotions (sadness, happiness, anger, and anxiety) and reported their empathetic experiences via brief questionnaires. Using data from these sensors and questionnaires, machine learning models were developed to predict empathy scores based on physiological responses during the VR sessions [9].

2 Materials and Data Collection Process

2.1 Materials and Setup for Empathy Elicitation in VR

To elicit empathy, we immersed participants in a 360° and 3D virtual environment, as VR has proven more effective than methods like 2D videos, workshops, or text-based exercises [8, 13, 17, 20]. We used videos featuring actors expressing four emotions—happiness, sadness, anger, and anxiousness—without additional content to avoid confounding factors [2]. Recognizing the impact of understanding emotional context, an audio narrative version was also created, followed by a corresponding video (50-120 seconds). To ensure gender balance, we recorded videos with two male and two female actors. Five versions were developed: four with narratives (two male, two female) and one non-narrative, where all emotions are portrayed by all actors without accompanying narratives. The non-narrative version allows gradual transitions between emotions, making it suitable for participants of all linguistic backgrounds.

Additionally, a 2-minute forest video ("The Amsterdam Forest in Springtime") was included at the start to establish a relaxed baseline and a roller coaster video ("Official 360 POV - Yukon Striker - Canada's Wonderland") at the end to control for non-empathic arousal. Both videos were sourced from YouTube.

Participants completed trait empathy questionnaires (QCAE) [14] and, after each emotion-specific video, provided feedback

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.9356

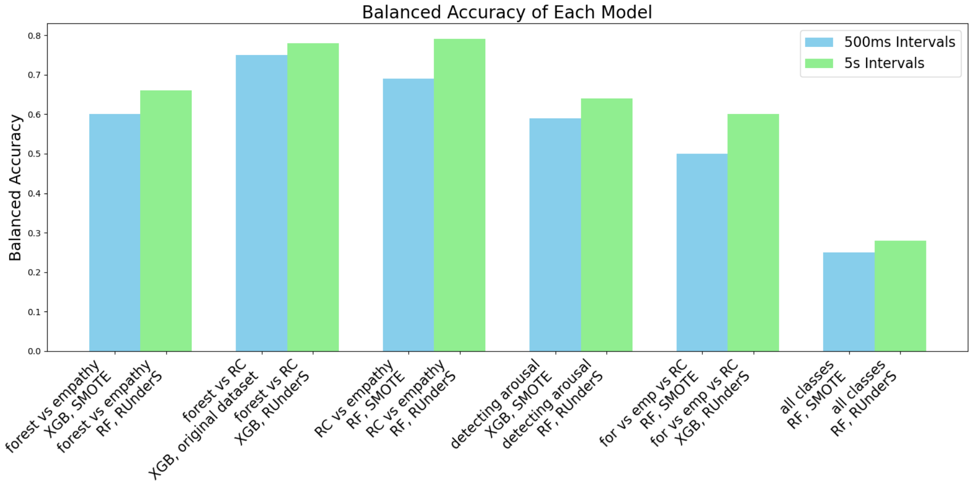

Figure 1: The best accuracies for each group of models, developed using datasets extracted at two different frequencies and various data balancing techniques, presented for all the labeling schemes

on their empathic state (State Empathy Scale) [18], arousal and valence levels (SAM) [3], and personal distress (IRI) [5]. Each VR session lasted around 20 minutes to minimize VR sickness, with participants viewing one of five versions.

Sensor data were collected using the emteqPRO system attached to the Pico Neo 3 Pro Eye VR headset, including EMG for facial muscle activation, PPG for heart rate, and IMU for head motion tracking. The device uses an internal clock as well [12].

2.2 Dataset Description

In this research, we used convenience sampling to recruit participants from the general public without a specific selection pattern. Participants were invited from various sources, including Jožef Stefan Institute employees, university students, and the general public. Invitations were sent verbally or in writing. Data collection concluded with 105 participants, averaging 22.43 ± 5.31 years (range 19–45), with 75.24% identifying as female. Participants had diverse educational and professional backgrounds. Additionally, ethical clearance for this study was obtained from the Research Ethics Committee at the Faculty of Arts, University of Maribor, Slovenia (No. 038-11-146/2023/13FFUM). Furthermore, written informed consent was obtained from the actors prior to recording.

The EmteqPRO system not only provides raw sensor data but also generates derived variables through the Emteq Emotion AI Engine, which utilizes data-fusion and machine learning to analyze multimodal sensor data and assess the user's emotional state. This system provides a file with 29 derived features, called affective insights for each recording: 7 features for heart-rate variability (HRV) and 3 for breathing rate; 2 features for facial expressions; 4 features for arousal and 4 for valence; 1 feature for facial activation; and 1 feature for facial valence. Additionally, head activity is tracked, reflecting the percentage of the session with head movement. Dry EMG electrodes on facial muscles such as the zygomatic, corrugator, frontalis, and orbicularis provide four more features, each representing muscle activation as a percentage of maximum activation observed during calibration. The data also includes the time elapsed since the start of the recording and the row index.

3 Methodology

3.1 Preprocessing

Since all the features or insights are numeric, except for the feature "Expression/Type," which has three values—smile, frown, and neutral—we applied one-hot encoding, a technique used in data preprocessing where categorical (non-numeric) variables are transformed into a numerical format. Each unique value in the original non-numeric feature is transformed into a separate binary (0 or 1) feature.

Next, because missing values represent less than 1% of the total data for each participant, they were filled in using the average of each feature's values. Scaling the values in the descriptive features between 0 and 1 was the final step in the preprocessing process.

3.2 Feature Engineering

Since features were provided at intervals ranging from 1 second to 500 milliseconds, we divided the data into two windows: one of 5 seconds and one of 500 milliseconds. For each window, we computed features from the 22 insights across the seven modules, as well as from the features for head activity and facial muscle electrodes, deriving a total of 108 new features, including minimum, maximum, average, and standard deviation for each original feature or insight. Additionally, the features for head activity and facial muscle electrodes were used to define 'Expression/Type,' and the time and row index were used as provided. However, the row index was disregarded further in the study.

We labeled the dataset in six different ways: 1) as a binary classification aiming to detect empathic arousal, comparing empathic parts with the forest part of the video, while excluding the non-empathic content of the roller coaster video; 2) as a binary classification using the forest and roller coaster, aiming to detect non-empathic arousal; 3) again, as a binary classification, but including only empathic parts and the roller coaster, aiming to distinguish between empathic and non-empathic arousal, and examining the differences in physiological responses between empathic content and non-empathic arousal-inducing content, such as the roller coaster video; 4) aiming to detect arousal in general, regardless of whether it is empathic or non-empathic, by splitting the entire dataset into two classes: the forest and everything else, including empathic parts and the roller coaster; 5) into three classes: treating the chunks of the roller coaster and forest as separate classes and grouping all the empathic parts into one class, without differentiating between the different emotions. The goal is to distinguish among no-arousal, empathic arousal, and non-empathic arousal; 6) with the average of participants' answers to the state empathy questions for each part of the video, with each part of the empathic content considered a separate chunk. Additionally, there are two other classes: the forest and the roller coaster. The aim is to detect the level of empathy participants experience during the session. We also included each participant's ID, intending to later use it for model evaluation with the 'leave-one-subject-out' technique.

4 Experiments and Results

4.1 Experimental setup

To build models for predicting a participant's state empathy during the VR session, we used six different classification algorithms: Gaussian Naive Bayes, Stochastic Gradient Descent Classifier, K-Nearest Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, and Extreme Gradient Boosting Classifier. The balanced accuracy score was used as an evaluation metric to assess the classification models for predicting participants' state empathy. This metric evaluates the overall balanced accuracy of the model by calculating the average of recall obtained on each class. Additionally, we used a confusion matrices to evaluate the performance of the classification models by comparing the actual and predicted labels.

For model evaluation, we used a Leave-One-Subject-Out cross-validation setup, where each subject is a unique participant identified by their ID.

Because the labeling schemes 2, 3, 5, and 6 are not balanced (with the 80% of the majority class), we conducted four experiments for each developed model: 1) applying the Synthetic Minority Over-sampling Technique (SMOTE) to create synthetic samples for the minority class to balance the dataset; 2) using the RandomUnderSampler (RUnderS) method to randomly select samples from the majority class, thereby reducing their count and balancing the dataset; 3) using SMOTETomek, a combination of SMOTE for oversampling and Tomek links for undersampling, which targets both the minority and majority classes; and 4) using the dataset as it is, without any undersampling or oversampling.

4.2 Results

Including models developed by six different classification algorithms on two distinct datasets—with two different window sizes—and utilizing four different data balancing techniques: undersampling, oversampling, combination techniques, and the dataset in its original form, along with six different labeling schemes, we obtained 288 unique confusion matrices and corresponding accuracies for each combination.

We ran a correlation matrix, which revealed that the highest correlation with the state empathy feature was found with the derived maximum and minimum values from the mean heart rate, the derived maximum and minimum values from the arousal class feature, and the average of the arousal class — the insight, which can be -1 (low), 0 (medium), or 1 (high). The derived standard deviation, maximum, and minimum values from the activation—expressed as a percentage of the maximum activation of particular muscles from the calibration session, especially the zygomaticus and orbicularis muscles—were also highly correlated.

Regarding the labeling schemes, we can conclude the following: 1) We can detect empathic arousal with confusion matrices that show a relatively good distribution of correct predictions across both classes and high accuracies for most of the developed models; 2) We can detect non-empathic arousal, with almost every developed model achieving a balanced accuracy higher than 60%, reaching up to 78%, and a reasonable balance between classes, indicating satisfactory classification performance; 3) We can even distinguish between empathic and non-empathic arousal with balanced accuracy of 79%; 4) We can detect arousal in general, again with high accuracies and balanced classes; 5) We can distinguish to some extent among no-arousal, empathic arousal, and non-empathic arousal; 6) However, it is currently very challenging to detect the precise level of empathy participants are feeling during the session using these methods, and to determine whether they are empathizing by mirroring emotions or experiencing something different while observing specific emotions. The best we can detect in this regard is up to 28% balanced accuracy, with confusion matrices showing a relatively balanced performance across multiple classes, with a good number of correct classifications, particularly in the more frequent classes.

Regarding the two window sizes, both models showed similar class balance and balanced accuracy scores. However, the dataset extracted at 5-second intervals performed slightly better. Using this dataset, false positives and false negatives were reduced more effectively. This led to more reliable classification performance, especially in terms of precision and recall, despite the smaller scale. Thus, the models developed using the 5-second interval dataset generally performed better, showing more effective classification and fewer errors. The simpler confusion matrix and potentially better handling of fewer classes suggest that it performs better in practical terms (Figure 2, Figure 1).

Regarding the data balancing techniques, the undersampling technique never produced the best results. For the dataset extracted at 500 ms intervals, using the SMOTE oversampling technique and SMOTETomek yielded the best results. For the dataset extracted at 5-second intervals, using the entire dataset yielded the best results, although models developed using SMOTETomek yielded slightly lower results in each combination of different labeling schemes.

Regarding the classification algorithms, Gaussian Naive Bayes performed the worst in terms of balanced confusion matrices, while Random Forest Classifier and Extreme Gradient Boosting performed the best across all combinations, with Random Forest Classifier showing slightly better results for most combinations (Figure 2, Figure 1).

4.3 Conclusion

In this study, we define the entire plan for developing materials, methods, and environments to evoke and measure the level of empathy. We started by defining the videos and the session, creating or selecting questionnaires for later use as ground truth, writing the narratives, recording the VR videos, and then editing and preparing them for use. Additionally, we collected a dataset from over 100 participants, which we filtered, preprocessed, and prepared for feature engineering and analysis.

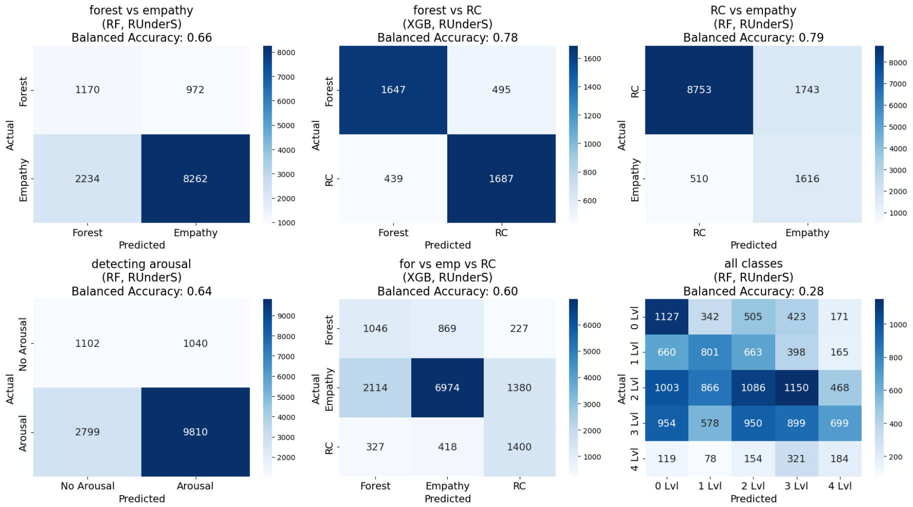

Figure 2: The best confusion matrices for each group of models, developed using dataset extracted at a 5s window size and various data balancing techniques, shown for all labeling schemes

We conducted and analysed four groups of experiments, totaling 288 combinations, where we developed models using two different window sizes, six classification algorithms, and three resampling techniques, with six different labeling schemes aimed at detecting various aspects of the dataset chunks: four empathetic parts, forest, and roller coaster.

The main conclusion is that we can detect arousal in general, non-empathic arousal, empathy, and differentiate between non-empathic and empathic arousal, as well as between relaxed states and arousal. However, we face difficulties in detecting and distinguishing between the precise levels of empathy during VR sessions using these methods and approaches.

Our next steps involve refining the detection of empathy levels during VR sessions by applying detailed data filtering and transforming it into a stationary format. Furthermore, we will develop models such as Autoregressive, Moving Average, and Extended Recurrent Moving Average, and use clustering techniques like DBSCAN and HDBSCAN. Additionally, we will extract more features from the raw data or use end-to-end neural networks. We plan to analyze gender differences in empathy with a t-test [7], and explore the impact of narrative context and emotions on empathic responses using ANOVA and MANOVA [4, 6].

Acknowledgements

The part of Emilija Kizhevska was supported by the Slovenian Research and Innovation Agency (ARIS) as part of the young researcher PhD program, grant PR-12879. The technical aspects of the videos, the recording and video editing were skillfully conducted by Igor Djilas and Luka Komar. The actors featured in the videos were Sara Janašković, Kristýna Šajtošová, Domen Puš, and Jure Žavbi. The questionnaires were selected and created, the narratives were written, and the psychological aspects of the video creation were considered by Kristina Šparemblek.

References

[1] D. Banakou, P. D. Hanumanthu, and M. Slater. 2016. Virtual embodiment of white people in a black virtual body leads to a sustained reduction in their implicit racial bias. *Frontiers in Human Neuroscience*, 10, 226766.

[2] P. Bertrand, J. Guegan, L. Robieux, C. A. McCall, and F. Zenasni. 2018. Learning empathy through virtual reality: multiple strategies for training empathy-related abilities using body ownership illusions in embodied virtual reality. *Frontiers in Robotics and AI*, 5, 326671.

[3] M. M. Bradley and P. J. Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25, 1, 49–59.

[4] A. Cuevas, M. Febrero, and R. Fraiman. 2004. An anova test for functional data. *Computational Statistics and Data Analysis*, 47, 1, 111–122.

[5] M. H. Davis. 1980. A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology/American Psychological Association*, 85.

[6] A. French, M. Macedo, J. Poulsen, T. Waterson, and A. Yu. 2008. Multivariate analysis of variance (manova). *San Francisco State University*.

[7] T. K. Kim. 2015. T test as a parametric statistic. *Korean Journal of Anesthesiology*, 68, 6, 540–546.

[8] E. Kizhevska, F. Ferreira-Brito, T. Guerreiro, and M. Luštrek. 2022. Using virtual reality to elicit empathy: a narrative review. *VR4Health@ MUM*, 19–22.

[9] E. Kizhevska, K. Šparemblek, and M. Luštrek. 2024. Protocol of the study for predicting empathy during vr sessions using sensor data and machine learning. *PloS One*, 19, 7, e0307385.

[10] F. F. D. Lima and F. D. L. Osório. 2011. Empathy: assessment instruments and psychometric quality–a systematic literature review with a meta-analysis of the past ten years. *Frontiers in Psychology*, 12. 781346.

[11] M. Mado, F. Herrera, K. Nowak, and J. Bailenson. 2021. Effect of virtual reality perspective-taking on related and unrelated contexts. *Cyberpsychology, Behavior, and Social Networking*, 24, 12, 839–845.

[12] M. J. Magnée, B. De Gelder, H. Van Engeland, and C. Kemner. 2007. Facial electromyographic responses to emotional information from faces and voices in individuals with pervasive developmental disorder. *Journal of Child Psychology and Psychiatry*, 48, 11, 1122–1130.

[13] K. M. Nelson, E. Anggraini, and A. Schlüter. 2020. Virtual reality as a tool for environmental conservation and fundraising. *PloS One*, 15, 4, e0223631.

[14] R. L. Reniers, R. Corcoran, R. Shryane Drake, N. M., and B. A. Völlm. 2011. The qcae: a questionnaire of cognitive and affective empathy. *Journal of personality assessment*, 93, 1, 84–95. DOI: doi:10.1080/00223891.2010.528484.

[15] G. Riva, J. A. Waterworth, and E. L. Waterworth. 2004. The layers of presence: a bio-cultural approach to understanding presence in natural and mediated environments. *CyberPsychology Behavior*, 7, 4, 402–416.

[16] R. O. Roswell, C. D. Cogburn, J. Tocco, J. Martinez, C. Bangeranye, J. N. Bailenson, and L. Smith. 2020. Cultivating empathy through virtual reality: advancing conversations about racism, inequity, and climate in medicine. *Academic Medicine*, 95, 12, 1882–1886.

[17] N. S. Schutte and E. J. Stilinović. 2017. Facilitating empathy through virtual reality. *Motivation and Emotion*, 41, 708–712.

[18] L. Shen. 2010. On a scale of state empathy during message processing. *Western Journal of Communication*, 74, 5, 504–524.

[19] M. Slater, A. Antley, A. Davison, D. Swapp, C. Guger, C. Barker, and M. V. Sanchez-Vives. 2006. A virtual reprise of the stanley milgram obedience experiments. *PloS One*, 1, 1, e39. DOI: 10.1145/1188913.1188915.

[20] J. Stargatt, S. Bhar, T. Petrovich, J. Bhowmik, D. Sykes, and K. Burns. 2021. The effects of virtual reality-based education on empathy and understanding of the physical environment for dementia care workers in australia: a controlled study. *Journal of Alzheimer's Disease*, 84, 3, 1247–1257.