Information Fusion 126 (2026) 103643

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Multi-modal physiological markers of arousal induced by CO_2 inhalation in Virtual Reality

Michal Gnacek [a,b,*], Neslihan Özhan [c], John Broulidakis [d], Ifigeneia Mavridou [e], Theodoros Kostoulas [f], Emili Balaguer-Ballester [g], Martin Gjoreski [h], Hristijan Gjoreski [i,b], Charles Nduka [b], Matthew Garner [c], Erich Graf [c], Ellen Seiss [j]

[a] Centre for Digital Entertainment, Faculty of Media and Communication, University of Bournemouth, Poole, BH12 5BB, United Kingdom
[b] Emteq Labs, Brighton, BN1 9RS, United Kingdom
[c] Department of Psychology, University of Southampton, Southampton, SO17 1BJ, United Kingdom
[d] Department of Experimental Psychology, University of Oxford, Oxford, OX1 2JD, United Kingdom
[e] Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, 5037 AB, Netherlands
[f] Department of Information and Communication Systems Engineering, University of the Aegean, Samos, Karlovasi 832 00, Greece
[g] Department of Computing and Informatics, Faculty of Science and Technology, Interdisciplinary Neuroscience Research Centre, University of Bournemouth, Poole, BH12 5BB, United Kingdom
[h] Faculty of Computer Science, Università della Svizzera italiana, Lugano, 6900, Switzerland
[i] Faculty of Electrical Engineering and Information Technology, Ss. Cyril and Methodius University, Skopje, 1000, North Macedonia
[j] Department of Psychology, Faculty of Science and Technology, Interdisciplinary Neuroscience Research Centre, University of Bournemouth, Poole, BH12 5BB, United Kingdom

ARTICLE INFO

Dataset link: www.gnacek.com

Keywords:
Affective computing
Physiological signals
Virtual reality
Methods of data collection
Arousal

ABSTRACT

High arousal states, like fear and anxiety, play a crucial role in organisms' adaptive responses to threats. Yet, inducing and reliably measuring such states within controlled settings presents challenges. This study uses a novel approach of CO_2 enriched air vs normal air in a Virtual Reality (VR) context to induce high arousal whilst measuring physiological signals such as galvanic skin response (GSR), facial skin impedance, facial electromyography (fEMG), photoplethysmography (PPG), breathing, and pupillometry. In a single-blind study, 63 participants underwent a regimen involving 20 min of breathing regular air followed by 20 min of 7.5% CO_2, separated by a brief interval. Findings demonstrate the efficacy of CO_2 inhalation in eliciting high arousal, as substantiated by statistically significant changes for all signals, further validated through high (94%) accuracy arousal classification. This study establishes a method for inducing high arousal states within a laboratory context validated through comprehensive multi-sensor data and machine learning analyses. The study underscores the value of employing a suite of physiological measures to comprehensively describe the intricate dynamics of arousal. The generated database is a promising resource for researching physiological markers of arousal, panic, fear, and anxiety, offering insights that can potentially resonate within clinical and therapeutic realms.

1. Introduction

Emotions are essential components of human experience which significantly impact our behaviour, cognition, and overall well-being [1]. Consequently, detecting and measuring affective states has been a topic of growing interest in psychology, neuroscience, and computer science [2]. Affective computing (AC), the field focused on developing algorithms and systems to recognise, interpret, and respond to human emotions, has gained particular attention in recent years [3]. AC can potentially improve many areas of human–computer interaction, including adaptive systems, gaming, and healthcare [4,5].

The reliable induction of affective states is a prerequisite for AC, leading to the co-existence of multiple emotion elicitation methods currently [6]. Typical forms of robust affective state inductions include the use of passive stimuli such as images [7], video [8] or sound [9]. The emotional responses to these states can be measured with arousal and valence self-ratings or physiological measures. Valence refers to the

* Corresponding author at: Centre for Digital Entertainment, Faculty of Media and Communication, University of Bournemouth, Poole, BH12 5BB, United Kingdom.
E-mail address: mgnacek@bournemouth.ac.uk (M. Gnacek).
URL: http://www.gnacek.com (M. Gnacek).

https://doi.org/10.1016/j.inffus.2025.103643
Received 4 May 2024; Received in revised form 18 July 2025; Accepted 20 August 2025
Available online 2 September 2025

pleasantness or unpleasantness of an emotion [10], while arousal reflects the level of activation or intensity of the emotional response [11].

However, a common problem in existing affective databases is that stimuli often fail to elicit extreme ends of the arousal spectrum [12,13]. In particular, very high arousal states, linked with threat processing, can be difficult to reproduce repeatedly and ethically in a laboratory setting [14]. Thus, models based on limited sections of the affect spectrum provide an incomplete picture. One solution could be the usage of immersive, interactive and engaging alternatives like VR (Virtual Reality) [15,16] typically leading to stronger emotional responses [14,17,18], which however can face an additional challenge — the metabolic demand. Interactive experiences such as VR by nature, involve a level of physicality and movement. This hampers the discrimination between, e.g., heart rate increases resulting from physical exertion or increasing arousal levels [19]. The same logic applies to other physiological affect indicators such as sweating (GSR) and respiration rates. Lastly, physical movement in interactive experiences often results in increased noise levels, potentially reducing data quality, and resulting in low reliability [20].

An alternative approach to induce enhanced arousal levels is the inhalation of CO_2 enriched air. The anxiogenic properties of CO_2 inhalation have enabled researchers in medical and pharmacological fields to study panic [21] and anxiety [22] disorders. Autonomic arousal changes triggered in response to inhalation of CO_2 enriched air (gas mixture with an increased amount of CO_2) [23] also make this method appealing to AC researchers wishing to induce and investigate arousal. This is particularly true when high arousal levels are desired because the severity and modulation of the reaction are reliable and quickly reversible [22].

Thus, we propose that the induction of high arousal in VR (via CO_2 inhalation whilst wearing a VR headset) can overcome subjectivity and cross-subject reliability issues and offer a more robust research method for affective dataset creation. To this end, we present here a study where 63 participants inhaled 20 min of regular atmospheric air (control) and 20 min of 7.5% CO_2 (21% O_2, nitrogen to fill) through an oronasal mask in two separate, single-blind conditions separated by a fifteen-minute break. We simultaneously recorded facial skin impedance and electromyography (EMG), photoplethysmography (PPG), pupil dilation and inertial measurement unit (IMU) physiological signals with a biofeedback-enabled emteqPro VR headset [24]; incorporating separate finger sensors for galvanic skin responses (GSR) and a breathing belt for registering respiratory movements. Below, we summarise our main contributions:

1. **CO_2-Based Physiological Signal Database:** to the best of our knowledge, this is one of the few existing physiological signal databases that utilise CO_2 inhalation, an approach that remains rare in the field.
2. **Multi-Modal Physiological Dataset:** data containing a rich array of physiological signals such as facial EMG, PPG, IMU, GSR, pupillometry and respiration in response to CO_2 exposure available for future research.
3. **Characterisation of Arousal Signals:** detailed analysis, including temporal information for all physiological measures collected in response to CO_2 inhalation (high arousal state)
4. **Wide Arousal Spectrum Coverage:** demonstrated a full spectrum of arousal states, from low (air condition) to high (CO_2 condition), enabling robust analysis and model training across the entire arousal range.
5. **Identified Most Suitable Physiological Predictors of Arousal**: Evaluated multiple physiological signals to determine the most informative predictors of arousal across both continuous and post-processed analyses.

2. Related work

This section encompasses the two key aspects coalescing in this study: the affect detection and induction methods within the context of AC, and the CO_2 enriched air inhalation process.

2.1. Affect detection

In the context of the three pillars of AC (emotion recognition, emotion expression and subjective experiences/feelings) [3], affect detection is a critical component and a pre-requisite for any affective system to respond to the user's mental state.

Typically, the subjective experience of an emotion is recorded with self-reporting measures, such as questionnaires or subjective rating scales. They allow individuals to directly express their emotional experiences (feelings) [25]. With this subjective experience, many physiological bodily changes co-occur and can be captured via physiological signals such as heart rate, skin impedance [26], and facial electromyography (EMG) [27]. Facial expressions [28], vocal prosody [29], and body language [30] are other cues that can be analysed using computer vision and/or audio processing techniques. Analysing these biosignals and using outputs to guide the system in recognising, interpreting and ultimately responding to human emotion is the goal of AC [31].

Machine learning algorithms play a vital role in affect detection by enabling the development of models that can learn patterns and relationships in the collected data [32]. Supervised learning approaches on labelled datasets are commonly used to recognise affect by combining feature extraction techniques and classification algorithms [33]. Classifiers ranging from classic support vector machines, random forests and shallow neural networks to deep classifiers can effectively categorise the input data into different affective states (see, for instance, [34]). More recently, ML affect detection approaches started incorporating multi-modal data fusion by combining multiple modalities like facial expressions, vocal cues, and physiological signals to improve the accuracy and reliability of emotion recognition [35,36].

2.2. Affect induction

The development of accurate affect detection algorithms relies on correct affective labels for physiological signals. Thus, reliable and reproducible affect induction is a prerequisite for developing affect detection systems. Common approaches to affect induction include the use of visual stimuli such as images [37] or videos [38], auditory stimuli [9], VR environments [39], memory recall [40], and even pharmacological interventions [41]. However, a systematic challenge in designing reliable affect induction is that certain induction methods favour specific emotions [6,42]. This fact, combined with the ongoing effort to enhance affect detection reliability, has led to the continual development of affective databases, which aim to leverage evolving technologies and target the entire affective spectrum [5].

Many affective databases include self-reported ratings to validate the participants' experience and to label data. Thus, they are prone to the subjective bias of participants' responses; this subjectivity issue underlies all media-based affective datasets. As a result, affective databases require large samples to validate their repeatability, and even then, it cannot be guaranteed when transitioning between different cultures and demographics [43,44]. Moreover, the distinction between the conscious experience of emotions (self-ratings), physiological measures of the body and brain during this experience, and facial/bodily expression of emotions exacerbate subjectivity issues of utilising self-ratings for affect detection [45].

These challenges suggest that an ideal affect induction method would have the following characteristics: (i) be fully reversible without any long-lasting side effects, (ii) reproducible for every individual, (iii) able to induce a full spectrum of affect and (iv) not rely on subjective self-ratings as labels.

Table 1
Study summary.

Number of participants	63 total-17 excluded, 46 used for analysis		
Conditions	1. Atmospheric air 2. 7.5% CO_2		
Condition duration	20 min each, 15 min break in-between		
Study design	Within-subjects		
Questionnaires	PHQ9, GAD, SPIN, MINI-7 (addiction and anxiety sections)		
Rating	Post condition: arousal/valence(1–9), anxiety(0–100)		
Physiological Signal	Device used	Frequency	Location
EMG	emteqPro	1000 Hz	7 channel–face
Facial skin impedance	emteqPro	50 Hz	7 channel–face
PPG	emteqPro	50 Hz	Forehead
Gyroscope	emteqPro	50 Hz	Head
Accelerometer	emteqPro	50 Hz	Head
GSR	Biopac	1000 Hz	Finger
Respiration	Biopac	1000 Hz	Chest
Eye tracking	HTC Vive Pro Eye (Tobii)	120 Hz	Eyes

2.3. CO_2 inhalation

The respiratory acidosis state is caused by the accumulation of carbon dioxide in the body and its impact on arousal has been a subject of interest in numerous studies [46]. In experimental settings, controlled exposure to CO_2 inhalation can induce acute anxiety and autonomic arousal in healthy individuals [47], impairing prefrontal executive functions such as cognitive performance [48] and threat processing [23]. This approach can be advantageous over regular induction methods due to its ability to intensify arousal symptoms and evoke negative affect in a dose-dependent manner, proving effective in eliciting a broad spectrum of responses in all individuals [49]. Moreover, CO_2 inhalation has been identified to reliably raise systolic blood pressure and heart rate, increase respiration frequency, and induce a sense of breathlessness, indicating heightened autonomic arousal [21]. Critical for this study, these physiological changes are always accompanied by subjective and initially dose-dependent subjective experiences of anxiety, fear, and stress [22]. By systematically manipulating the CO_2 concentration and exposure duration, researchers can elicit targeted arousal levels for affective states, and affect-driven technology design. This novel method for affect induction, initially validated in pharmaceutical studies for testing anxiolytic medications, fulfils the desired criteria by being completely reversible, consistent across participants, and not reliant on subjective self-ratings.

Carbon dioxide is generally regarded as having low toxicity when inhaled, but it can still pose significant risks, particularly in specialised occupations or controlled environments [50]. While the safety of CO_2 inhalation in research settings has been extensively studied and validated across numerous publications [21,51,52], this does not imply that experiments can be conducted with unrestricted concentrations or durations. In most studies involving CO_2 inhalation, stringent exclusion criteria are typically applied, and the interplay between inhalation duration and CO_2 concentration is a critical factor in ensuring participant safety. The likelihood of side effects from CO_2 inhalation increases with both the duration of exposure and the concentration level, effectively reducing the maximum safe exposure time for an individual [50]. At one end of the spectrum, some studies have employed extremely high CO_2 concentrations up to 35% but limited exposure to a single breath [53], which limits the duration and amount of physiological data that can be collected during such a short period. At the other end of the spectrum, CO_2 concentrations below 1% typically produce little to no observable effects [50]. Researchers select specific concentrations based on experimental goals with higher concentrations such as 20% or 35%, delivered through single or double vital capacity inhalations, which are commonly used to elicit acute panic responses, while prolonged exposures lasting up to 20 min require lower concentrations typically between 5% and 7% to induce states of anxiety, fear, tension, and stress [49]. As a result, a 7.5% CO_2 concentration was selected for this study to maximise the duration of exposure, allowing for the collection of extensive physiological data while maintaining participant safety and eliciting sufficiently strong effects.

3. Experimental setup

Healthy participants recruited from the general public participated in two separate, single-blind conditions (air and CO_2 inhalation) using a within-subject design. Each condition lasted 20 min and was preceded by sensor fitting and calibration. Participants were seated for the entire duration. An “air condition” was administered first. A gas hissing sound was continuously played from a speaker throughout both conditions.

Participants were not aware of the active condition. After each condition and during a 15 min long break, participants took off all equipment, and proceeded to a separate preparation room where they were offered water and were asked to provide their arousal and valence ratings (self-assessment manikin 1–9), anxiety ratings (slider 0–100), and to predict the condition (air or CO_2). Fig. 1 shows an overview of the study protocol while 5 illustrates the experimental setup combining all equipment, hardware and software.

3.1. Recruitment and participants

Ethics approval was obtained from the University of Southampton (71104.A2). Initial screening exclusion criteria from [23] were used, and several additional exclusions were added. The complete prescreening list included age (under 18 or over 55), body mass index (under 18 or above 28), weight under 45 kg, pregnancy — suspected or confirmed, breastfeeding, history of panic disorders (participant or family), diagnosed cardiovascular, respiratory, neurological conditions, diabetes, severe allergies, current or recent participant in another medical trial, acute illness in the past 7 days, any medication use in past 8 weeks (not including paracetamol, aspirin, topical treatment, contraceptives), history of migraines requiring treatment, self-reported history of alcohol/drug abuse, regular smokers (1 or more cigarette per day), COVID-19 diagnosis within the last month or LONG COVID-19, not being registered with a GP (General Practitioner/local healthcare provider) and finally, having a large beard as it interfered with oronasal mask seal.

On the day of the data collection, each participant had their blood pressure tested to ensure it was within the predetermined limit (less than 140/90 with resting heart rate below 90 beats per minute). A breathalyser test was also administered to ensure participants had not consumed any alcohol before the study. A large proportion of female participants was unable to obtain a satisfactory oronasal mask seal, and female recruitment was stopped after initial testing.

In total, 63 participants took part in the study of which 59 were male and 4 were female. Eight participants asked to stop the CO_2 condition before it was complete and were excluded from the analysis. Nine additional participants were excluded due to data corruption or equipment malfunction. The remaining 46 participants consisted of 45 males and 1 female with a mean age of 23.48 years (min = 18, max = 41, SD = 5.830).

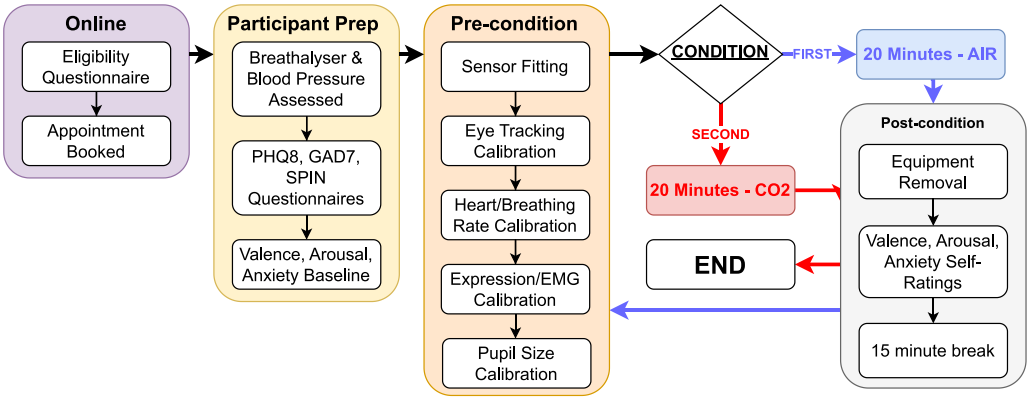


Fig. 1. Flow diagram illustrating the study protocol.

3.2. Hardware

The air was supplied to participants through a face-worn oronasal mask (RESMED AirFit F30 [54]). A custom adapter was 3D printed to connect the oronasal mask to a tube connected to the air reservoir bag. This bag contained a switch that allowed to change between air sources for the two conditions (regular atmospheric and CO_2 enriched air). The reservoir bag was connected to a regulator attached to medically certified gas mixture cylinders (7.5% CO_2, 21% O_2, N_2 balance, sourced from a local supplier). Cylinders were securely fastened to a wall. The regulator allowed for pressure and flow rate regulation. To compensate for changes in gas consumption due to varied respiratory rates between and within participants over time, the flow rate was adjusted by one of the supervising researchers as needed to ensure the reservoir bag was full and the participant had a large supply of air available at any time.

The emteqPro system [24] mounted on a HTC Vive Pro Eye headset [55] enabled us to collect physiological measures, including forehead PPG, fEMG (7 channels - centre corrugator and left/right for zygomaticus, frontalis and orbicularis muscles), facial skin impedance (same 7 channels), head movements (gyroscope and accelerometer). Eye tracking data were recorded using the Tobii eye tracking sensors embedded within the HTC VIVE PRO headset [56] (see Fig. 2). BIOPAC MP150 was used to include GSR (GSR100C) and respiration belt (RSP100C) sensor models [57].

3.3. Software

We developed a custom VR application using the Unity engine integrating emteqPro SDK and Tobii eye tracking package. The developed app integrated step-by-step instructions for each calibration step, recorded events of interest, and simultaneously triggered data collection from emteqPRO and eye sensors. The VR scene for both conditions was an empty, grey, dimly-lit room with uniform lighting. It contained a couch situated below the user's point of view to convey the impression of sitting. Participants were instructed to look straight ahead in a virtual environment that was deliberately devoid of visual points of interest. Aside from the ability to look around, no interactions, movements, or animations were included. This minimalistic scene was designed to simulate the experience of sitting in an empty room, providing a controlled and distraction-free setting in contrast to the actual laboratory environment, which contained various sensors, computers, and medical equipment.

The SuperVision application for the emteqPro device (v1.4.0) and the BIOPAC application (AcqKnowledge version 4.3) were used for real-time signal monitoring [24].

3.4. Data processing

Fig. 3 depicts the data-processing approach. A Windows 10 machine (GTX1060 6 GB, i7-6700 and 32 GB of RAM) with a shared system clock was the host system for the study, enabling time synchronisation of signals from different devices. All data were initially up-scaled to match the highest signal frequency of 1000 Hz using forward filling. Once synchronised, it was then down-sampled to 50 Hz. Data were divided into four different types of files. EmteqPro generated two of the four files, namely '.dab' file containing physiological signals from the device. These dab files were converted into '.csv' files using software provided by the manufacturer for loading into a Python environment for further processing. Secondly, '.json' files contained event information (custom message indicating start/end of calibration and condition segments).

A similar process was used for processing and storing pupil dilation data in a proprietary '.eyedata' format from Tobii. Data were initially converted to readable 'csv'. The pupil dilation processing pipeline included outlier rejection steps as described in [58] and included specific techniques for the following outlier types: (1) invalid data outliers (provided by the eye tracking device, e.g., due to eye being closed, momentary glitch etc.), (2) feasible range outliers (e.g., pupil dilation was outside the predetermined, feasible range of 1.5–9 mm), (3) dilation speed outliers (e.g., pupil dilation change was disproportionately large to adjacent samples), and (4) gap artefacts outliers (e.g., pupil dilation was invalid from all of the combined filtering steps described). These samples are also likely invalid due to eye occlusion by blinking or other artefacts. Thus, we removed the frame corresponding to 120 hz before and after each gap. Given the high correlation between both pupils [59], in the absence of pupil dilation from one eye, valid pupil dilation from the other eye replaced missing data. Once the remaining outliers have been identified and removed, we used mean linear interpolation to fill in gaps in pupil dilation data. Finally, timestamps enabled us to synchronise pupil dilation data with the emteqPro and event data.

The last data file considered was a readable '.csv' file containing GSR and respiration data from the BIOPAC device. In contrast to all other data, this file included the recording's start time instead of timestamps. Given a known frequency of 1000 Hz, we used the start time to generate timestamps for each data row.

Signals from all devices were synchronised and downsampled to 50 Hz, followed by a 3rd order Butterworth low-pass filter of pupil dilation, GSR, PPG, contact and IMU (Gyroscope and Acceleration) signals and a 4th order on respiration data. Filtering was unnecessary for the EMG data since the manufacturer already filtered and processed the *EMG Amplitude* for all channels. Lastly, using data from both conditions and calibration segments, each participant's dataset was individually normalised using min–max normalisation.

Fig. 2. Researcher demonstrating emteqPro VR device with the oronasal mask attached to a hose via a custom 3D printed adapter.

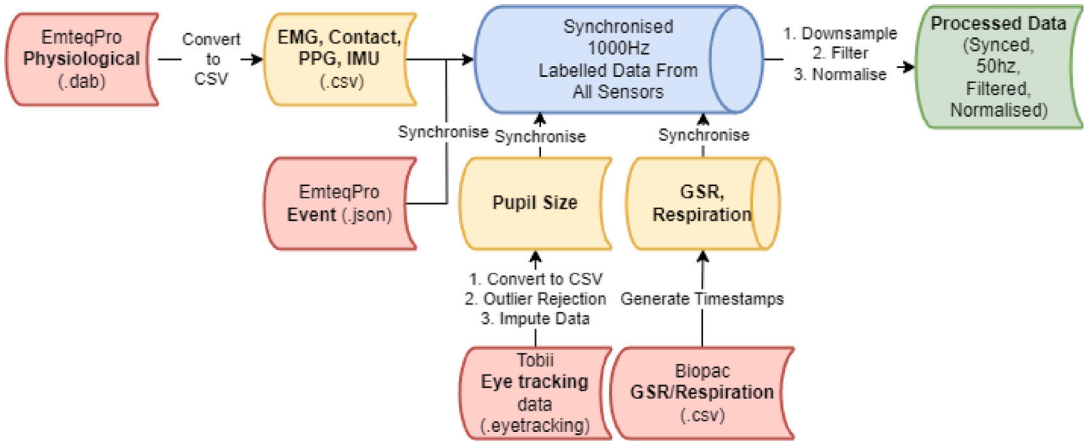


Fig. 3. Flow diagram depicting the data processing pipeline.

3.5. Feature extraction

All signals (refer to Table 1) had the following features extracted: mean, std, min, max, median, range, iqr, 1st and 2nd derivative means and std. Combined mean pupil dilation across both eyes was also calculated using left and right-eye pupil dilation.

In addition, the neurokit2 library allowed us to extract additional, signal-specific features from the heart-rate sensors: PPG_Rate_Mean (beats per minute), HRV_MeanNN, HRV_SDNN, HRV_RMSSD, HRV_SD SD, HRV_MedianNN, HRV_IQRNN for PPG data, SCR_Peaks_N, SCR_Peaks_Amplitude_Mean, EDA_Tonic_SD for GSR data, RSP_Rate_Mean (breaths per minute), RSP_Amplitude_Mean and RSP_Phase_Duration _Ratio for respiration data.

Feature extraction was performed separately for segment overview and timed windows. In the segment overview, features were extracted from the entire duration of each condition, i.e., one mean heart rate value for the CO_2 gas inhalation condition and one for air condition per participant. Additional feature extraction was performed on separate time windows (each 60-s long and 10-s slide) of signals generated from each segment, resulting in, e.g., 115 mean heart-rate values extracted per condition for each participant.

Fig. 4 heatmap shows Pearson correlation coefficients between time-windowed means of physiological signals, indicating that features from separate sensors are predominantly uncorrelated. Most moderate to strong correlations exist between features extracted from the same modality, like heart rate magnitude and variability, EMG data from pairs of muscle sensors, and left and right pupil dilation data. A salient exception is the respiration rate, which shows moderate to strong correlations with several other features such as heart rate, heart rate variability, GSR and certain EMG features.

4. Data analysis

This section investigates the effect of the CO_2 inhalation on participant physiological responses compared to baseline air condition. First, we analysed each modality individually to identify differences between

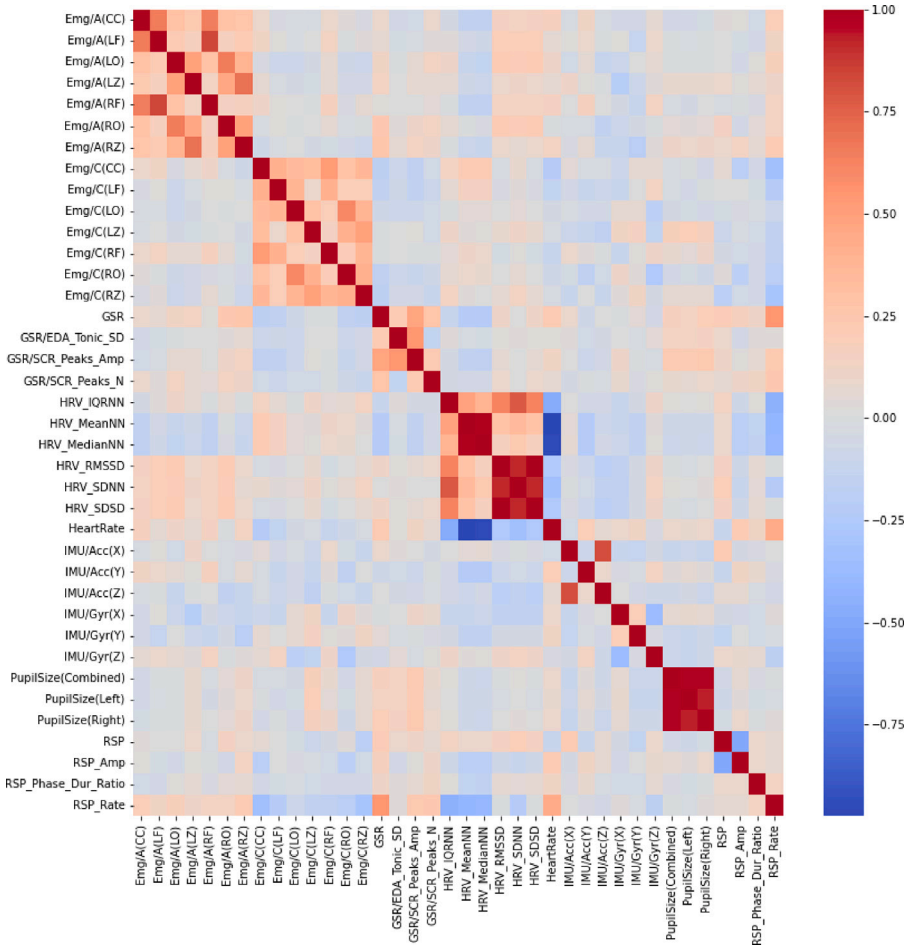


Fig. 4. Pearson correlation coefficients heat map for a subset of extracted features.

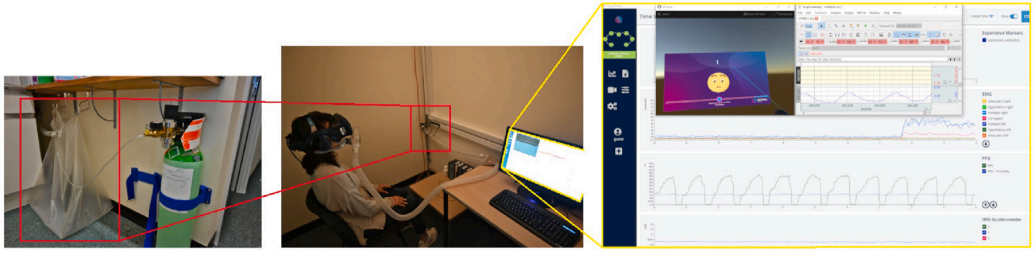


Fig. 5. Experimental setup. From left to right: (i) Cylinder connected to a reservoir bag; (ii) researcher wearing VR headset, physiological sensors and oronasal mask; (iii) supervising researcher view of the real-time signals through SuperVision and AcqKnowledge applications.

air and CO_2 conditions temporal dynamics. To this end, Fig. 6 shows the mean values of GSR, pupil dilatation, respiration and heart rates time series and overall violin plots.

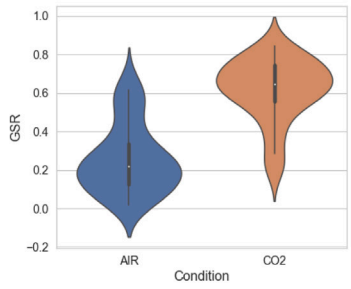
4.1. Condition validation with self-ratings

The arousal, valence (1–9) and anxiety ratings (0–100) before the start of the experiment and after each condition were analysed to evaluate the effect of condition (air vs. CO_2) on the ratings (see Fig. 7).

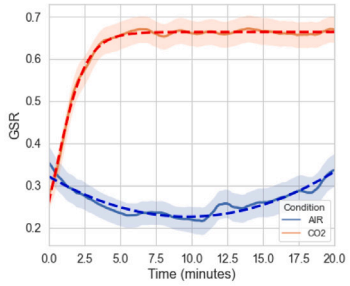
Starting with arousal ratings, a repeated measures one-way ANOVA (within-participant factor of time - baseline/post air/post-CO_2) with a Greenhouse–Geisser correction determined that mean arousal ratings differed significantly between conditions ($F(1.696, 71.239) = 66.614$, $p < 0.001$, $\eta^2 = 0.613$). Bonferroni corrected pairwise comparisons revealed that arousal ratings were not significantly different between baseline and air condition ($\bar{d} = 0.302$, $se = 0.259$, $p = 0.749$) but were significantly different between baseline and CO_2 ($\bar{d} = -3.001$, $se = 0.374$, $p < 0.001$) and between air and CO_2 ($\bar{d} = -3.302$, $se = 0.306$, $p < 0.001$).

Likewise, this approach revealed significant differences in valence ratings between all three conditions ($F(1.389, 59.717) = 87.897$, $p < 0.001$, $\eta^2 = 0.671$). Post-hoc analyses revealed significantly different valence ratings for all pairwise comparisons: baseline and air condition ($\bar{d} = -0.705$, $se = 0.154$, $p < 0.001$), baseline and CO_2 ($\bar{d} = -3.227$, $se = 0.312$, $p < 0.001$), air and CO_2 ($\bar{d} = -2.523$, $se = 0.275$ $p < 0.001$).

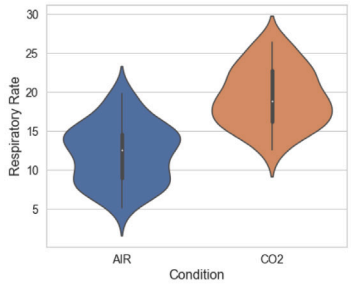
Finally, anxiety ratings differed between conditions ($F(1.236, 43.247) = 61.808$, $p < 0.001$, $\eta^2 = 0.638$). Post-hoc analyses revealed anxiety ratings shared a similar pattern to arousal, with anxiety ratings not differing between baseline and air conditions ($\bar{d} = 1.708$, $se = 1.717$, $p = 0.980$), but reaching high significance for baseline and CO_2 ($\bar{d} = -32.714$, $se = 3.817$, $p < 0.001$) as well as air and CO_2 conditions ($\bar{d} = -34.422$, $se = 4.360$, $p < 0.001$).



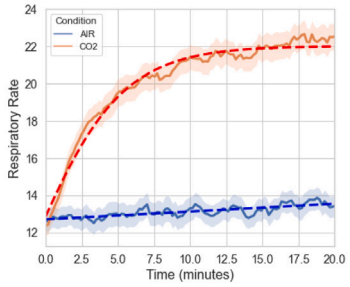
(a) Galvanic skin response distributions.



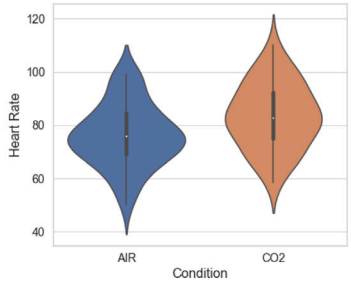
(b) Time series mean values of galvanic skin responses.



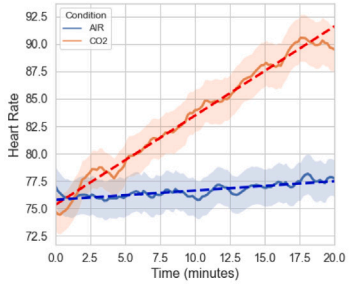
(c) Respiration rates distributions.



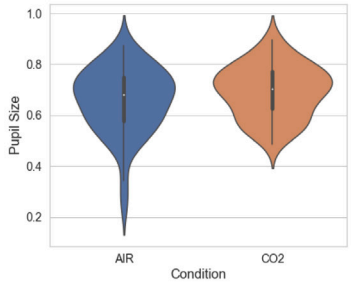
(d) Time series mean values of respiration rates.



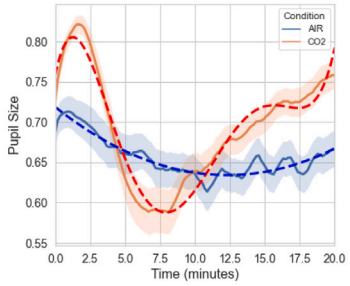
(e) Heart rates distributions.



(f) Time series mean values of heart rates.



(g) Pupil dilation data distributions.



(h) Time series mean values of pupil dilation data.

Fig. 6. Air and CO_2 enriched air conditions comparing GSR, respiration rate, heart rate and pupil dilation measures across participants. Left violin plots (a, c, e, g) show the distribution of values for each measure and condition. Right line plots (b, d, f, h) display the time series mean values for 60-s (with a 10-s overlap) sliding time windows per condition.

4.2. Galvanic skin response

The mean values of GSR for the air and CO_2 conditions differ significantly (paired t-test $t(45) = 8.538, p < 0.001, d = 2.120$). Fig. 6 displays the time courses of mean GSR levels for the air and CO_2 conditions. The CO_2-related GSR dynamics are effectively captured by a logistic function curve fitting, revealing a sigmoidal response pattern characterised by maximum GSR, growth rate, and temporal midpoint parameters. Meanwhile, a quadratic function successfully explains the air-related GSR dynamics, portraying a parabolic trend with parameters reflecting amplitude, curvature, and temporal shift.

4.3. Respiratory rate

Along similar lines, mean respiration rates (RR) between the air and CO_2 conditions differ significantly ($t(45) = 14.216, p < 0.001, d = 1.913$).

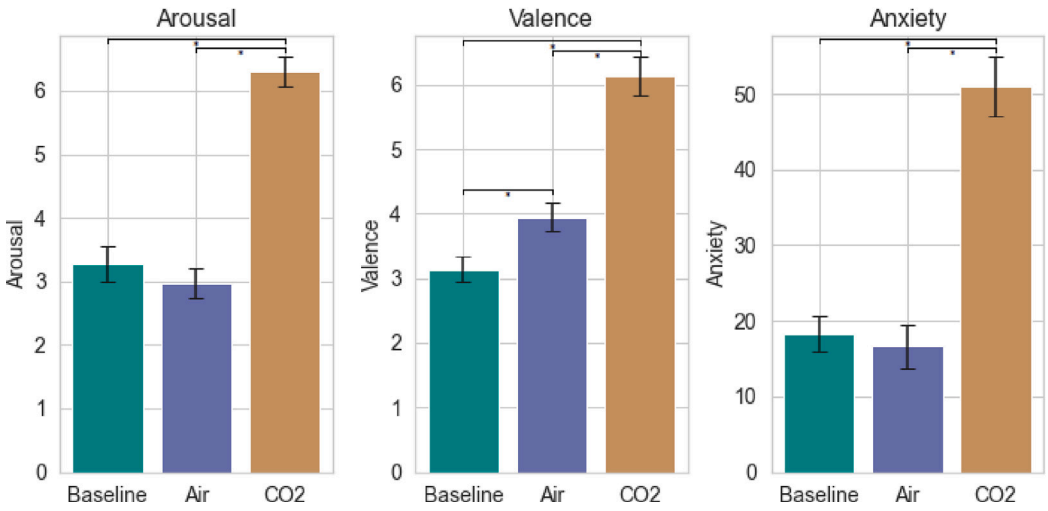


Fig. 7. Mean and standard error of arousal (1–9), valence (1–9) and anxiety (0–100) ratings provided before the study (baseline), and after both air and CO_2 conditions (∗ indicates significance at $p < 0.001$, see main text).

Temporal dynamics under elevated CO_2 levels follow a sigmoidal response pattern, while the RR dynamics associated with air exposure follow a linear trend (Fig. 6d).

4.4. Heart rate

The difference between the mean heart-rate values (beats per minute calculated from the emteqPRO PPG data) between the air and CO_2 conditions was also statistically significant (paired t-test, $t(45) = 45.533$, $p < 0.001$, $d = 9.490$). In contrast to respiration rate and GSR for the CO_2 condition, the temporal dynamics of heart rate exhibited a predominantly linear increase. This trend was also evident in the air condition, albeit with a considerably diminished slope compared to the CO_2 condition (Fig. 6f). Notably, the increase levels off subtly during the last few minutes of the CO_2 condition, hinting at a potential maximum of the heart rate.

4.5. Pupil dilation

Normalised pupil dilation did not differ significantly between conditions when comparing mean values independent of time ($t(45) = 1.52$, $p = 0.134$, $d = 0.309$). However, the picture differs when considering temporal dynamics, i.e., the pupil dilation in the air condition follows a predominantly quadratic model. Temporal changes in the CO_2 condition show a more complex behaviour modelled by a 5th-order polynomial (Fig. 6h). Initial spikes in pupil dilation are evident in both conditions. However, the CO_2 condition exhibits a higher and longer spike, followed by a rapid decline leading to a significantly lower minimum value than the air condition. After 6-8 min, the pupil dilation increases progressively more linearly.

4.6. EMG sensors: Facial muscle amplitude and skin impedance

The same sensors registered both EMG and skin impedance. The bar plots in Fig. 8 and time series plots in Fig. 9 provide a comprehensive visualisation of the mean values per condition across all time points and for each time window, respectively.

Next, we examined the effects of air and CO_2 conditions on facial EMG activity and skin impedance (Fig. 8). Individual muscle groups differed between the air and CO_2 conditions (paired t-tests, $p < 0.05$). Likewise, skin impedance measures significantly differ between conditions for all sensors apart from the "RightFrontalis". For the EMG amplitude measure, both sides of the orbicularis and zygomaticus muscles showed significantly different activation levels. Results are summarised in Table 2. For both measures, time series are displayed in Fig. 8. Curves were fitted with polynomial regressions for all conditions and sensors separately.

Table 2
Paired t-tests results for the skin impedance and EMG amplitude analyses.

Muscle	Skin impedance		EMG	
	T-statistic	*P*-value	T-statistic	*P*-value
LeftFrontalis	2.921	0.005	−1.068	0.291
RightFrontalis	1.973	0.055	−1.100	0.277
LeftOrbicularis	2.085	0.043	−2.031	0.048
RightOrbicularis	3.512	0.001	−2.556	0.014
LeftZygomaticus	2.363	0.023	−2.747	0.009
RightZygomaticus	4.244	0.000	−4.794	0.000
CenterCorrugator	4.969	0.000	−1.427	0.160

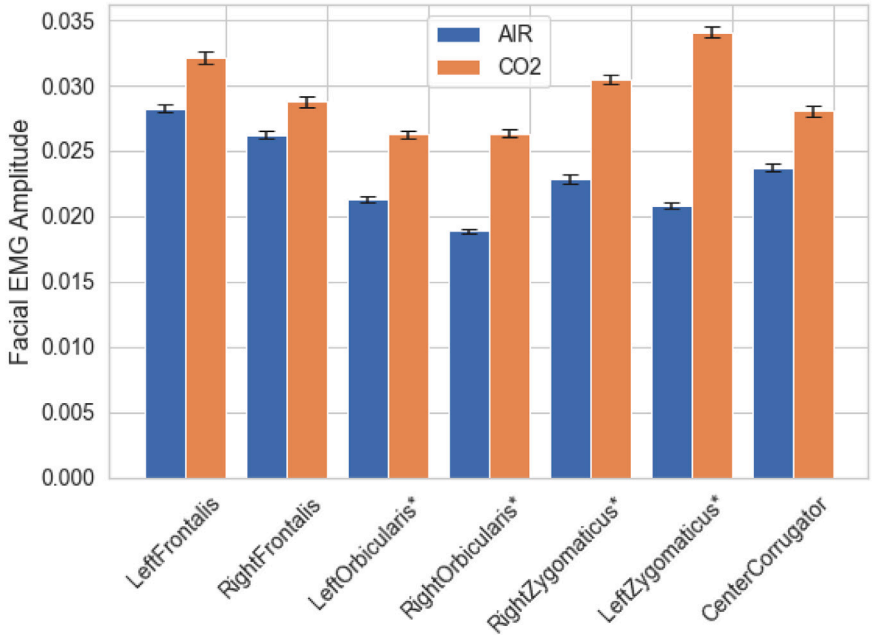
4.7. IMU (Gyroscope & Accelerometer)

In contrast with other physiological sensors, mean values for the IMU sensors used (gyroscope and accelerometer) were less informative (Fig. 10), as expected: Gyroscopes measure angular velocity, and accelerometers capture linear acceleration. Unless there is continuous movement, the signal from these sensors drops to its baseline. This characteristic, combined with our sedentary protocol, results in signals that contain short peaks which are not well represented by the arithmetic mean of the overall measurement. Therefore, other features previously mentioned, such as standard, deviation, interquartile range, min, max etc. were extracted instead.

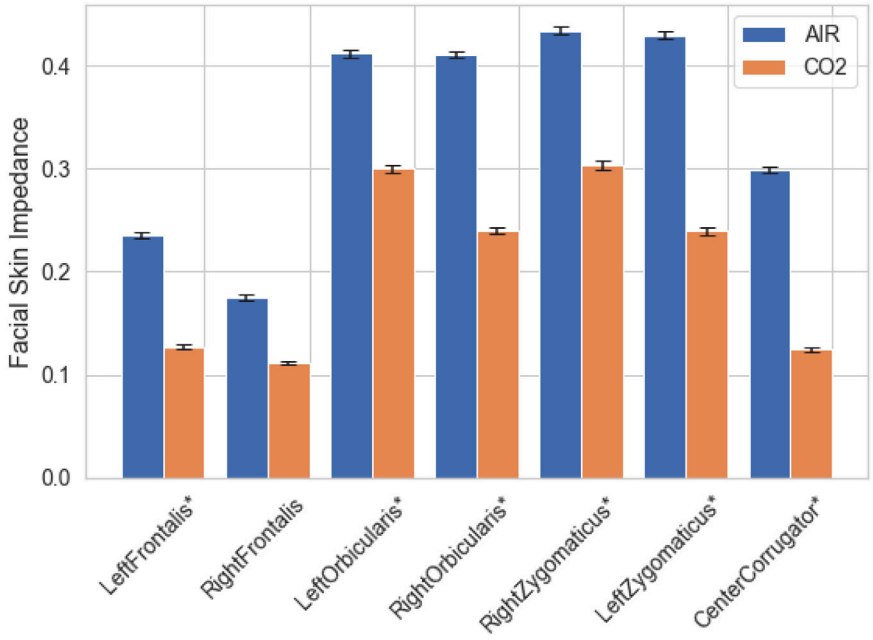
The combined results from a multivariate analysis of variance (MANOVA) for the gyroscope and accelerometer data revealed significant differences between conditions. Results underscore a substantial impact of both gyroscope features and condition on the observed variability (Wilks' $\Lambda = 0.004$, $F(28, 63) = 527.890$, $p < 0.001$). The condition itself (Air vs. CO_2) elicited a significant effect ($\Lambda = 0.362$, $F(30, 61) = 3.583$, $p < 0.001$), highlighting differences in gyroscope measurements between conditions.

Similarly, in the accelerometer analysis, the significance of the model was salient ($\Lambda = 0.0082$, $F(28, 63) = 272.371$, $p < 0.001$). The 'condition variable' remained statistically significant ($\Lambda = 0.415$, $F(30, 61) = 2.869$, $p = 0.0001$), further highlighting differences in accelerometer measurements.

MANOVA results provided robust evidence on the effect of condition in gyroscope and accelerometer variance. To explore the individual



(a) Overall EMG amplitude per condition.



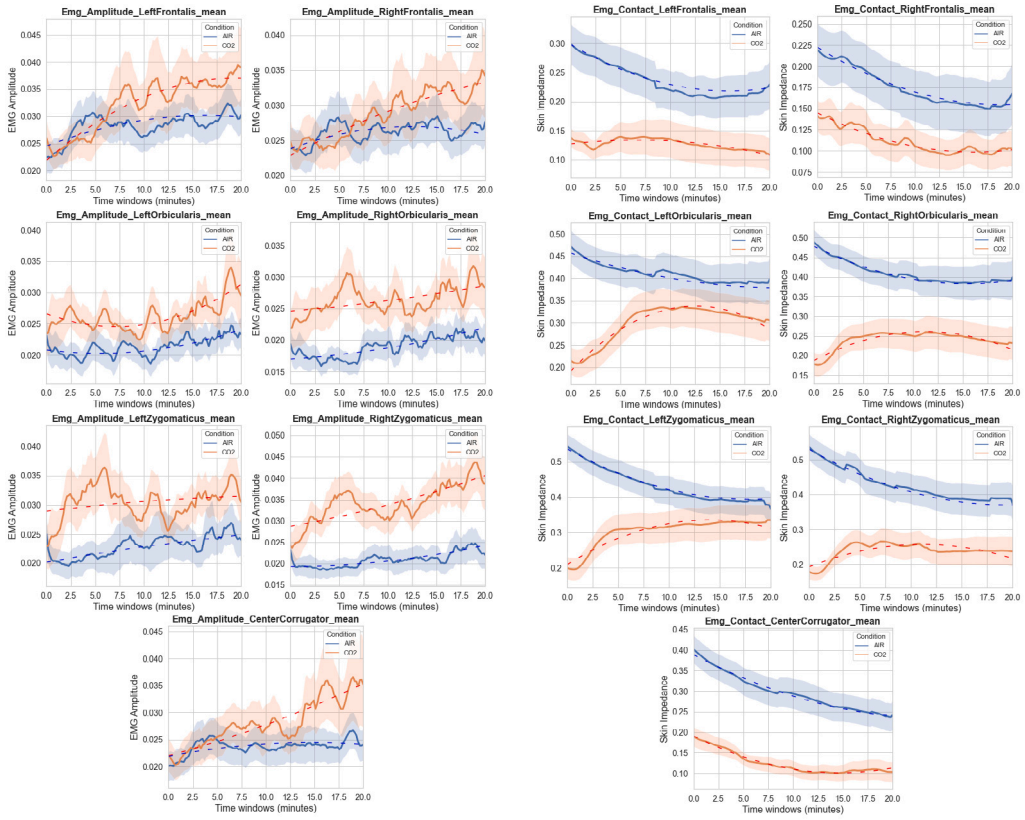
(b) Overall facial skin impedance per condition.

Fig. 8. Comparison of mean EMG amplitude and facial skin impedances between the air and the CO_2 conditions. $*$ indicates significance at $p < 0.001$. (a) Violin plots of mean values. (b) Regression fits.

impact of each feature, uni-variate ANOVA follow-up tests were carried out (full results table available in supplementary material). In summary, for the gyroscope, Bonferroni corrected univariate ANOVA analysis showed significant differences for interquartile range (IQR) as well as standard deviation (SD) and its two derivatives (first and second-order) across all three axes. For the accelerometer, the x-axis reached significance in SD, the y-axis reached significance in SD and the 1st derivative, and the z-axis for its first and second-order SD derivatives. The magnitudes of both sensors were calculated using the Euclidean norm (Fig. 10). The mean magnitude for the accelerometer also differs between conditions $t(90) = -3.098$, $p = 0.003$, but not for the gyroscope $t(90) = 0.002$, $p = 0.999$.

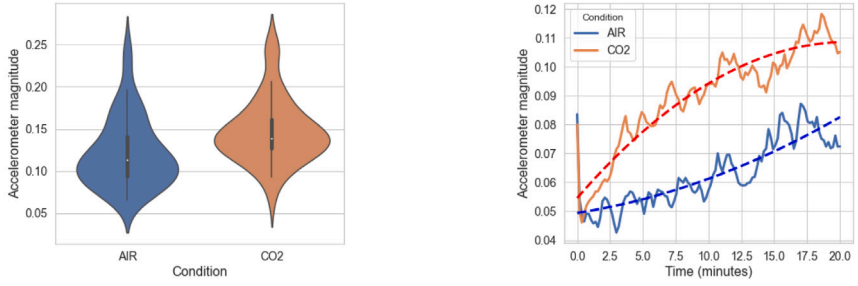
5. Machine learning decoding pipelines

We devised two distinct machine-learning pipelines, a feature-based and a deep learning-based pipeline. The first classic pipeline to assess feature differences between the air and CO_2 conditions. To this end, decoders utilised features extracted from physiological signals throughout their entire duration as inputs. The second and novel pipeline leveraged state-of-the-art deep learning algorithms to learn from raw

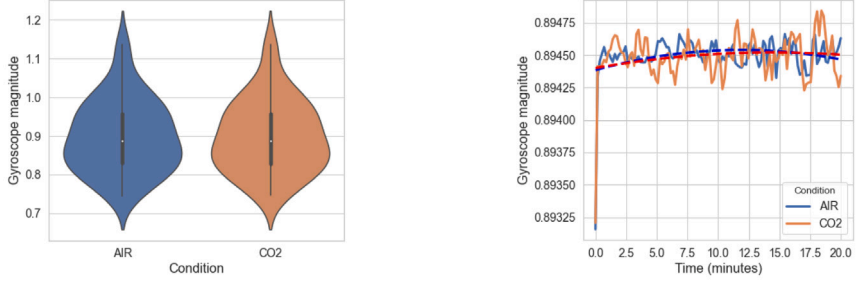


(a) Facial EMG amplitude over time. (b) Facial skin impedance over time.

Fig. 9. Comparison of mean EMG amplitude and facial skin impedances between air and CO_2 conditions over time for seven muscle groups with fitted regressions. (a) Regression fits for facial EMG amplitude. (b) Regression fits for facial skin impedances.



(a) Accelerometer magnitude distributions. (b) Time series of accelerometer mean magnitudes.



(c) Gyroscope magnitude distributions. (d) Time series of gyroscope mean magnitudes.

Fig. 10. Comparison between air and CO_2 conditions of probability densities and temporal dynamics of accelerometer and gyroscope mean values.

Table 3
Average accuracy (ACC) and F1-scores over 46 participants, grouped per data modality and classification method.

Single classification — Entire condition data								
Classifier	C-SVC		KNN		RF		RR	
modality	ACC	F1	ACC	F1	ACC	F1	ACC	F1
All	0.72	0.70	0.71	0.69	0.90	0.89	0.94	0.94
EMG A	0.57	0.50	0.61	0.60	0.66	0.62	0.62	0.59
EMG C	0.67	0.63	0.69	0.67	0.73	0.71	0.70	0.68
HRV	0.69	0.66	0.69	0.67	0.71	0.68	0.76	0.74
IMU	0.66	0.64	0.58	0.56	0.79	0.77	0.82	0.81
GSR	0.81	0.79	0.77	0.76	0.80	0.78	0.82	0.80
PupilSize	0.50	0.41	0.55	0.53	0.58	0.56	0.64	0.60
RSP	0.82	0.80	0.88	0.87	0.89	0.88	0.90	0.89
HRV/IMU/GSR/RSP	0.72	0.70	0.71	0.69	0.86	0.85	0.94	0.94
Continuous classification — Time windowed data								
Classifier	STResNet		CNN		ConvLSTM		Transformer	
modality	ACC	F1	ACC	F1	ACC	F1	ACC	F1
All	0.69	0.67	0.66	0.66	0.62	0.62	0.63	0.62
EMG A	0.69	0.69	0.64	0.63	0.63	0.62	0.58	0.57
EMG C	0.55	0.52	0.57	0.57	0.52	0.50	0.59	0.54
HRV	0.79	0.79	0.73	0.73	0.75	0.75	0.50	0.50
IMU	0.85	0.85	0.82	0.82	0.82	0.82	0.64	0.62
GSR	0.64	0.62	0.67	0.65	0.66	0.65	0.61	0.58
PupilSize	0.66	0.66	0.61	0.61	0.56	0.53	0.50	0.48
RSP	0.86	0.86	0.84	0.84	0.83	0.83	0.64	0.62
HRV/IMU/GSR/RSP	0.84	0.84	0.69	0.68	0.72	0.72	0.63	0.63

signal segments without previous feature extraction (30 s with 25 s overlap). Predictions generated in this pipeline were continuous (one prediction every 5 s), and thus is very different from the condition classification provided in the first pipeline. Thus, results are not directly comparable between the two approaches.

Table 3 lists accuracy and F1-scores of classification results for all decoders in both pipelines, as will be discussed next. Importantly, every model in both pipelines considers each input modality separately and the combination of all to demonstrate the differences of a multi-modal approach in arousal detection.

5.1. Feature-based analysis to distinguish between the air and CO_2 conditions

First, we used classic machine learning decoders to distinguish between the air and the CO_2 conditions based on the extracted physiological features. Baseline classifiers include C-SVC (Support Vector Classifier), KNN (K-Nearest Neighbours), RF (Random Forest), and LRR (Linear Ridge Regression), offering a range of distinct characteristics to tackle the classification problem from complementary angles, and were successfully used in affect detection [60].

In short, hyperparameter optimisation proceeded by defining a separate hyperparameter grid for each classifier, including regularisation strength, kernel type, number of neighbours, and maximum tree depth among others (see, for instance, [61]). This optimisation was followed by an independent leave-one-subject-out Cross-Validation (LOSO-CV) approach to ensure a reliable evaluation, preventing over-fitting. This technique involves iterative training classifiers on the data from all subjects except one and then evaluating the performance on the held-out subject. Repeating this process for each subject separately results in a comprehensive assessment of classifiers' generalisation capability. The heat map in Fig. 11 depicts classifier results for each out of the 46 participants. The x- and *y*-axis represent the test participant and input modality, while the colour shows the F1-score. Results show a high decoding accuracy for the multi-modal LRR decoder followed by the RF consistently for all participants, followed by RF, underpinned by the effect of respiration and GSR inputs primarily (Fig. 11).

5.2. Deep learning for continuous predictions with raw signals

Next, we devised a novel end-to-end continuous classification task for affect decoding leveraging raw signals. The diverse set of signals encompassed 7 EMG amplitude channels (EMG A), 7 EMG contact channels (EMG C), raw PPG and emteqPro-calculated heart rate (HR), the X, Y, and Z dimensions of the gyroscope and accelerometer instruments (IMU), right and left pupil dilation (PupilSize), Biopac galvanic skin response (GSR), and respiration rate (RR). Noticeably, like in the classic pipeline, we tested each modality individually (uni-modal approach) to gauge the accuracy of each signal independently, followed by a comprehensive multi-modal approach combining all input modalities.

To this end, we evaluated a range of state-of-the-art optimised Deep Learning architectures, including Convolutional Neural Networks (CNN - 3 layers, 64 units), Convolutional LSTM – Long Short-Term Memory networks (3 CNN and 1 LSTM layer, 64 units), a Transformer (2 layers, 2 heads, 16 units), and a STResNet [62] designed for multi-channel time-series data specifically; all with a consistent dropout rate of 0.2. Code implementation of each architecture is available with our published Python library at https://github.com/michalgnacek/co2-study.

Like in the previous section, we trained subject-independent models, i.e., the test subjects were not part of the training data. We grouped the subjects into five non-overlapping folds to speed up training and evaluation times and used a five-fold cross-validation strategy. Next, we split training data further into *internal* training and validation sets. Thus, for each fold, the data from the 46 subjects consists of training (32 subjects), validation (5 subjects) and test subjects (9 subjects). The validation data was used to monitor the model's performance during training on unseen subjects. The best-performing model on the validation data was used for the final test evaluation [62].

Models were trained using categorical cross-entropy loss and Adam optimiser (default learning rate of 0.001), and the training process was monitored using accuracy as the primary metric. Additionally, the supplementary material shows the training histories, including accuracy and loss curves over epochs to provide insights into the learning dynamics of each model. Even in this challenging end-to-end modality, the STResNet provides a reasonable overall continual accuracy of up to 87% for Respiration and 86% for IMU inputs, followed by the CNN architecture (3).

6. Discussion

This study investigated the effect of CO_2-enriched air vs. normal air on affective self-ratings. We leveraged physiological and movement measures to discern whether supplemented air inhalation can induce affective, physiological, and movement changes over time in a VR environment. This question is relevant because affect/mood induction is usually achieved instead by affective stimuli such as pictures, words, text passages, music/videos, etc. Existing methods can induce salient valence differences [10]. However, the elicitation of arousal, especially in the realm of high values, presents a more complex challenge [18,24]. As alternative examples, fear [14] or stress-inducing public speaking [39] have been used to elevate arousal. However, the effectiveness of these methods can vary substantially among participants due to variable interpersonal traits [44] and movement noise.

By contrast, CO_2-enriched air inhalation is a dependable and safe approach for inducing heightened arousal states, presenting multiple advantages over conventional methodologies [23]. This study introduces an innovative approach for collecting self-ratings, GSR, EMG, respiration rate, heart rate, and head movement, including acceleration and angular velocity. In addition, recording facial EMG presented an extra challenge due to CO_2 inhalation necessitating an oronasal mask obscuring the entire face. The new EmteqPro device addressed this challenge with a new monitoring approach of muscle activity from dry facial EMG sensors (and a PPG sensor) integrated into a VR HMD

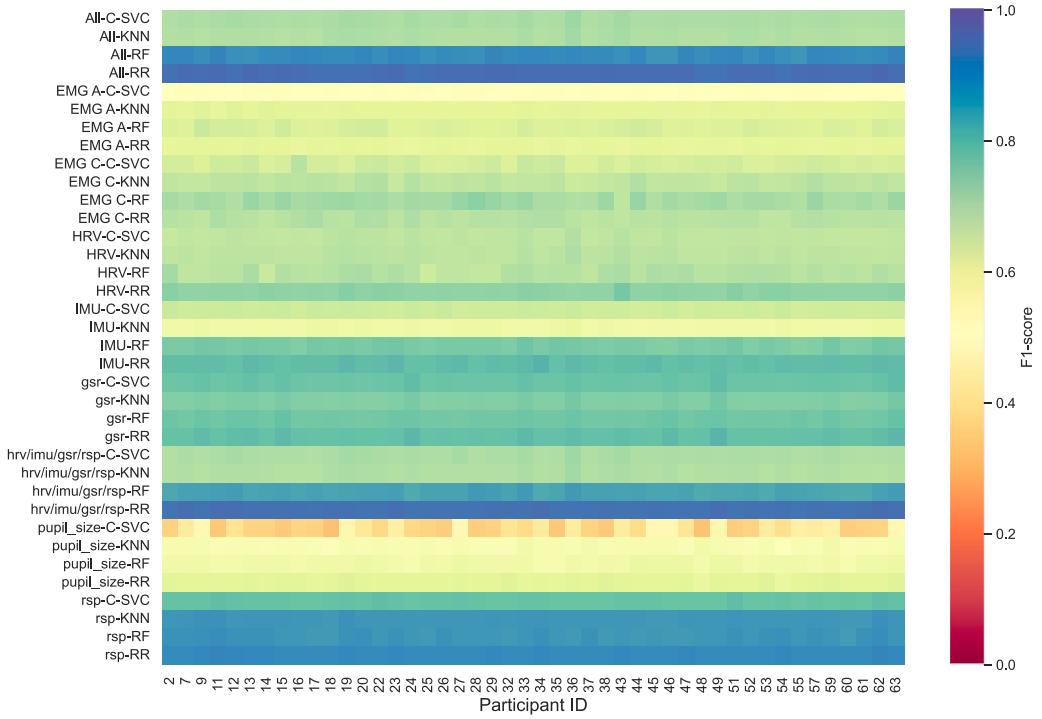


Fig. 11. Heat map showing F1-scores for individual participants per classifier and modality in air vs. CO_2 classification task.

headset. To our knowledge, this study is the first to combine facial EMG recordings with CO_2 inhalation, not to mention its integration into a VR setup, which could enable future studies to display virtual environments in conjunction with CO_2 inhalation.

Remarkably, our findings showed that the CO_2-enriched air condition induced higher arousal and anxiety levels than the regular air condition, reflected in all physiological and movement measures. Multimodal physio-facial data also permits the development of feature-based, end-to-end shallow and deep learning models with generally high test accuracy (Table 3) that could serve as baselines for developing more advanced solutions for monitoring arousal changes.

6.1. Self-ratings, participant experience and experimenter observations

We registered arousal, valence and anxiety self-rating levels at the beginning and end of each condition. As expected, all participants experienced the CO_2 condition as more arousing and anxiety-provoking than the baseline (air) condition (Fig. 7). A common theme for all participants was a pattern of strenuous and pronounced breathing that underwent an initial sharp escalation, followed by a subsequent tapering off and eventual stabilisation at a consistent pace. This experience was frequently characterised as fraught with stress, worry and apprehension, which often peaks during the first minutes of the experiment, followed by the acclimatisation to the heightened state. Self-regulation techniques like controlled breathing or positive cognitive and emotion regulation strategies facilitated this adjustment. These reported experiences align with the quantitative self-ratings and hence are central to interpreting temporal patterns of physiological changes underlying arousal dynamics.

6.2. Physiological and movement analysis

Most physiological measures showed significant differences between the CO_2 and the air condition. Specifically, mean respiration rate, heart rate, galvanic skin response (GSR), and facial EMG amplitudes were enhanced while skin impedance levels diminished in the CO_2 condition. All measures indicated a heightened state of arousal in the CO_2 in line with the collected self-ratings and previous research [23], except the mean pupil dilation, which did not differ between the conditions (see details in the following section).

More interestingly, temporal patterns of physiological signals were manifold. Physiological responses generally (but see below) showed a consistent escalation, followed by a levelling off – a phenomenon indicative of a potential threshold – in response to CO_2 inhalation, consistent with participants' experiences. However, their temporal dynamics were diverse.

Specifically, the GSR signal reached its plateau much more swiftly than other metrics (between 2.5 and 5 min). Previous research exposing participants to the same amount of CO_2 albeit for a much shorter duration (2 min) found a similar pattern of GSR, rising sharply and followed by a steady level [49]. The respiration rate sharply increased for the first 10 min, continuing to climb gradually until approximately the 15 min mark. This profile will yield a high discrimination power between air and CO_2 conditions discussed in the next section. The availability of precise data regarding the temporal shifts in respiration rate during CO_2 inhalation is limited (emphasising the significance of this paper). Still, a heightened breathing rate is a well-recognised symptom of acidosis [46].

Interestingly, the heart rate measure did not follow this pattern but exhibited a near-linear increase almost up to the end of the condition. While other CO_2 studies primarily present mean heart rate values for a general condition [21,23], consistent with our findings, the present study accentuates temporal differences between respiratory and cardiovascular control systems, elucidating how these interactions can lead to unique response patterns during CO_2 inhalation.

Crucially, this study unveils heightened activation of the EMG amplitudes measured by the facial sensors across all facial muscles in the CO_2 condition. The simultaneous facial muscle activation across the facial muscles may signify arousal intensity changes, whereas changes in specific muscles (orbicularis and zygomaticus) might relate more to valence changes [27]. However, we observed increased deglutition (swallowing) behaviour for most participants during the CO_2 condition, possibly influencing EMG amplitude in these areas. This study did not measure primary facial muscle activity involved in swallowing.

Additionally, the proximity of the buccinator (one of these muscles) to the zygomaticus muscle may have some impact [63].

The skin impedance from the facial EMG sensors allowed us to evaluate if separate GSR measurements outside the EmteqPro mask were necessary, gauging the information provided by impedance responses. Skin impedance signals typically decrease over time (Fig. 9). However, responses increased for the oribularis and zygomaticus sensors (reduced contact/higher impedance) before decreasing again (enhanced contact/lower impedance) only in the CO_2 condition, presented last. Additionally, the timing of the peaks of skin impedance approximately aligns with those of GSR. This timing may suggest success in capturing the relationship between GSR and skin impedance. Amplitude plots (Fig. 9) offer an alternative explanation for this phenomenon. The presence of EMG amplitude peaks indicate heightened facial movements during these segments in the CO_2 condition. This presence could explain skin impedance increase due to worsened skin contact caused by facial movements.

Tasks involving hyperventilation (such as those due to CO_2 inhalation) are typically characterised by elevated levels of head motion that synchronise with the inhalation and exhalation cycles [64]. The temporal analysis of the accelerometer revealed a consistent increase in magnitude over time for both conditions, with a noticeably higher trend in the CO_2 condition. The baseline air measurements suggested that participants might have been gradually disengaging or feeling discomfort due to prolonged stillness and the presence of numerous sensors. Accordingly, the CO_2 condition led to a significantly higher and more pronounced increase of the accelerometer magnitude over time.

Finally, pupil dilation analyses led to intriguing findings. The use of a VR environment prevented pupil dilation due to luminosity changes. In the CO_2 condition, the time course analysis revealed a swiftest and very pronounced dilation followed by a prominent pupil contraction, culminating in a slower, sustained, and linear phase of pupil dilation. In the air condition, pupil dilation partially correlated with the GSR signal, forming an inverse U-shaped trajectory (Fig. 6).

Pupils generally dilate in response to increases in arousal [58]. The effect of deliberate inhalation of CO_2 and pupil size has not been well researched, and existing literature is scarce and often limited to individual case reports of unconscious patients [65]. However, it suggests that increased concentration of CO_2 levels in the blood leads to pupil contraction [66], in line with our findings. The interplay between possibly opposing effects of CO_2 inhalation, arousal, and increased breathing rate on CO_2 bloodstream concentration and potential self-regulation may account for the observed pupil dilation curve with multiple inflexion points.

6.3. Effective decoding arousal levels

Machine learning classifiers effectively discerned between high (CO_2) and low arousal air conditions from physiological signals. Multimodal classifiers based on all input features achieved a very high accuracy of up to 94% for the LRR decoder and a comparable F1 score (Table 3), not previously achieved to our knowledge. Respiration features garnered the highest accuracy, followed closely by IMU and GSR consistently for all subjects (Fig. 11), underlying the multimodal approach success. We expected the relevance of these inputs, given the sharp and consistent increase of such physiological responses from the outset of the CO_2 stimulation (Fig. 6). By contrast, the EMG amplitude, conventionally used for valence detection, demonstrated the least accuracy among the variables [67] again in line with its more variable temporal dynamics (Fig. 9). Furthermore, we developed a continuous prediction end-to-end deep learning pipeline to delve into the interaction between physiological measures and arousal. Surprisingly, several single-input models surpassed the comprehensive approach incorporating all available modalities (Table 3). This decrease in accuracy for the multi-modal model, which combined all signals, can be attributed to the dilution effect of fusing informative with less relevant signals [68], suggesting specific input selection strategy might be more beneficial. Based on this finding, we combined the four best-performing modalities, including HRV, IMU, GSR, and respiration to evaluate the impact of excluding lower-performing modalities on classification accuracy and repeated the experiment for both pipelines. Taking a step back to the single classification pipeline where the multimodal approach already yielded the highest accuracy within the single classification pipeline, our results showed that adding pupil size, EMG amplitude, and contact data offered little to no improvement. This suggests that these modalities may be safely omitted, potentially enhancing performance by reducing the number of features evaluated. We conducted similar experiments using the continuous classification pipeline, evaluating models that used only the four modalities (HRV, IMU, GSR, and respiration). These models achieved higher accuracy than those trained on the full set of input signals. However, the best results overall were still obtained by the top-performing single-modality models.

Notably, the STResNet classifier based on the respiration rate signal exhibited an accuracy of 0.87, followed closely by IMU with 0.86, and HRV with 0.81. It is worth highlighting a decline in the performance of GSR between the two approaches. This observation suggests that despite often being regarded as a gold standard for arousal detection [14,34], GSR might not be optimal for continuous detection within shorter time windows. Alternative measures such as respiration rate, IMU and heart rate could offer further insights in such scenarios.

Respiration rate consistently achieved the highest classification accuracy across both single and continuous classification pipelines (see Table 3), likely due to its strong and direct physiological response to CO_2 inhalation. Elevated CO_2 levels trigger an immediate and automatic increase in respiration rate as the body attempts to regulate [49]. This strong, rapid, and sustained response made respiration rate particularly well-suited for classification, offering a robust signal with clearly distinguishable features. Unlike other physiological measures that may plateau or vary in latency, respiration rate continues to rise over time, supporting effective continuous classification throughout the exposure period (see Fig. 6).

6.4. Limitations and future work

This study enabled the simultaneous administration of CO_2 enriched air via the Resmed F30 oronasal mask and the recording of facial EMG with the EmteqPro mask integrated into the Vive Pro Eye headset. However, this setup featured some limitations. Even with this new compact oronasal mask, individuals with smaller facial proportions encountered fitting issues due to potential interference between the two devices, that is, an insufficient gap between the VR headset and the mask. This issue caused the headset to press and break the seal around the mouth, compromising gas delivery. Unfortunately, these participants had to be excluded, highlighting a unique hardware compatibility issue. Notably, this challenge disproportionately affected female participants, possibly introducing a bias. Future studies could explore options such as employing more compact VR headsets or tailored, alternative gas delivery mechanisms to mitigate this concern. For example, the HTC Vive Pro Eye headset is notably bulkier than other commercial headsets, e.g., Oculus Quest, and may not face the same problem.

Secondly, the CO_2 inhalation caused a heightened deglutition reflex and a dry mouth sensation. This effect could cause artefacts in the facial EMG measures unrelated to arousal changes. Future studies might find ways to extract these artefacts with ICA analysis techniques.

Previous studies showed that photoplethysmography (PPG) signals are a dependable source of respiratory rate measurements [69]. The database provided here furnishes a valid avenue for algorithm validation, offering both PPG and respiration belt data. Moreover, IMU sensors have the potential to enhance this algorithm through the fusion of PPG signals with subtle head movements that occur during deep

breathing. A sedentary,low-movement protocol can accentuate the subtle variations required for accurate respiratory rate interpretation from these signals, aiding the creation of more robust algorithms.

Another potential limitation of the study is the lack of counterbalancing in the order of experimental conditions. Specifically, all participants began with the air condition. In a preceding pilot study, the condition order was counterbalanced, with half of the participants starting with air and the other half with CO_2. However, the protocol was modified in the current study for two key reasons. First, the study employed a single-blind design in which participants were not informed of the condition they were undergoing. After each condition, participants were asked to guess which one involved CO_2. In every case, participants correctly identified the CO_2 condition, likely due to its unmistakable physical intensity. If the CO_2 condition were administered first, participants would easily infer that the following condition was air, thereby undermining the blind design. Second, the pilot study revealed that participants who began with the CO_2 condition were significantly more likely to withdraw from the experiment. While the precise reasons for this remain unclear, one plausible explanation is that the combination of multiple physiological sensors and a face mask created an environment that exacerbated discomfort or claustrophobic sensations. Starting with the air condition may have allowed participants to acclimate to the setup before undergoing the more challenging CO_2 condition. Based on these considerations, the decision was made to always begin with the air condition in the current study to preserve the data integrity and blind study design. We recommend that future studies investigate a counterbalanced design with a reduced level of CO_2 to reduce the likelihood of participant withdrawal.

An alternative and potentially valuable direction for future research using the collected data involves examining the role of anxiety, specifically how varying levels of anxiety may influence physiological responses. While our results confirmed that participants reported higher levels of anxiety during the CO_2 condition, a reverse analysis—investigating how baseline anxiety levels in the general population or even anxiety disorders modulates physiological signals was not conducted, as it fell outside the scope of the current study. Actually, participants who withdrew due to extreme panic or anxiety were excluded from the primary analysis. However, their data may hold important insights, as these individuals could exhibit the most pronounced physiological responses. Exploring their data in future work may help illuminate how their profiles differ from those who were able to complete the study.

The established reliability of CO_2 inhalation for inducing heightened arousal, corroborated by prior research [21,23,49] and our results, implies that overall condition ratings might be somewhat redundant beyond condition validation. However, our physiological findings indicate a non-linear and multi-directional pattern of arousal escalation (Figs. 9, 6). As a result, continuous self-ratings of arousal and other ratings could offer valuable insights across the condition scenario, fostering a more nuanced labelling system beyond the binary categorisation of high and low arousal.

Lastly, to enhance the interpretability of our machine learning models, future work should incorporate explainable AI (XAI) techniques [70]. For the feature-based classifiers, model-agnostic methods such as SHAP (SHapley Additive exPlanations) values and permutation importance can identify which physiological features most strongly influence model predictions. For the deep learning pipeline, methods such as integrated gradients, saliency maps, or attention-weight visualisations may shed light on which temporal segments or signal channels contribute most to continuous arousal predictions. Additionally, counterfactual explanations could offer valuable insight by showing how minimal changes in physiological features might alter the predicted arousal state, e.g., by identifying the threshold at which a respiration rate shift would flip a prediction from low to high arousal [71]. Integrating such XAI methods could also support the discovery of interpretable and potentially actionable physiological markers.

7. Conclusion

The present study proposes a novel approach to induce high arousal states through CO_2 inhalation while simultaneously capturing a range of physiological signals. The benefits of this study and the resulting database are multifaceted, offering valuable insights into affect induction and potential applications. While this approach is unlikely to be practical for everyday use due to the complexity of the setup and potential health risks, these extensive and laboratory-based experiments can yield valuable insights that support the development of more comprehensive arousal detection algorithms, the evaluation of emotion-aware systems, and the modelling of stress responses [72–74].

The challenges associated with reliably inducing high arousal, especially in static settings, are well-documented. This study successfully addresses this issue by demonstrating the efficacy of CO_2 inhalation in eliciting high arousal states. Through a combination of physiological measures such as respiration rate, heart rate, galvanic skin response, pupil size, and facial electromyography, the study showcases the distinctive responses induced by CO_2 inhalation. These responses, characterised by initial sharp increases followed by plateauing or subtle changes, provide a comprehensive view of the body's physiological reactions to high arousal induction by CO_2 inhalation.

Integrating the oronasal mask, the emteqPro mask and a breathing belt allow for simultaneous monitoring of facial EMG, movement, heart rate, eye tracking, and respiration rate measures. This fusion can lighten the interplay between facial expressions, physiological responses, and subjective affective states evoked by CO_2 versus air inhalation.

Machine learning decoders confirm the effectiveness of these combined physiological measures in discerning high arousal states. The high accuracy in distinguishing between CO_2 and air conditions – mainly through leveraging respiration rate dynamics – reinforces the approach's viability. Future work could simulate scenarios with less distinguishable affect conditions to explore the limits of the ML approach by systematically manipulating VR content.

Despite these contributions, the study acknowledges limitations such as the hardware fit constraints of the VR headset and the oronasal mask that potentially introduced gender bias due to the differing face sizes. Future studies may explore alternative solutions to mitigate these limitations and add continuous self-ratings of arousal to offer a more nuanced understanding of arousal dynamics.

In conclusion, this study's innovative approach of combining CO_2 versus air inhalation, VR, and comprehensive physiological sensing provides a new resource for researchers interested in investigating high arousal states. Hence, it contributes to a more holistic understanding of human affective experiences.

CRediT authorship contribution statement

Michal Gnacek: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Neslihan Özhan:** Resources, Project administration. **John Broulidakis:** Project administration, Methodology, Conceptualization. **Ifigeneia Mavridou:** Writing – review & editing, Formal analysis, Data curation, Conceptualization. **Theodoros Kostoulas:** Writing – original draft, Visualization, Formal analysis, Data curation. **Emili Balaguer-Ballester:** Writing – review & editing, Visualization, Validation, Methodology, Funding acquisition, Formal analysis. **Martin Gjoreski:** Writing – review & editing, Investigation, Formal analysis. **Hristijan Gjoreski:** Writing – review & editing, Investigation, Formal analysis. **Charles Nduka:** Supervision, Funding acquisition. **Matthew Garner:** Supervision, Resources, Project administration, Methodology. **Erich Graf:** Supervision, Resources, Project administration, Methodology. **Ellen Seiss:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Michal Gnacek reports financial support was provided by Emteq Labs. Charles Nduka reports a relationship with Emteq Labs that includes: board membership. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by Bournemouth University and Emteq ltd. via the Centre for Digital Entertainment (EPSRC Grant No. EP/L016540/1). Dr. M. Gjoreski's work was funded by SNSF, Switzerland through the project XAI-PAC (Grant No. PZ00P2_216405)

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.inffus.2025.103643.

Data availability

All data and code has been made available. Please visit www.gnacek.com for more information.

References

[1] T. Chambel, E. Oliveira, P. Martins, Being happy, healthy and whole watching movies that affect our emotions, in: Affective Computing and Intelligent Interaction, Vol. 6974 LNCS, Springer, Berlin, Heidelberg, 2011, pp. 35–45, http://dx.doi.org/10.1007/978-3-642-24600-5_7, URL https://link.springer.com/chapter/10.1007/978-3-642-24600-5_7.

[2] R. Corive, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J.G. Taylor, Emotion recognition in human-computer interaction, IEEE Signal Process. Mag. 18 (2001) 32–80, http://dx.doi.org/10.1109/79.911197.

[3] R.P. W., Affective Computing, The MIT Press, 2000.

[4] J. Tao, T. Tan, Affective computing: A review, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 3784 LNCS (2005) 981–995, http://dx.doi.org/10.1007/11573548_125/COVER, URL https://link.springer.com/chapter/10.1007/11573548_125.

[5] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, W. Zhang, A systematic review on affective computing: emotion models, databases, and recent advances, Inf. Fusion 83–84 (2022) 19–52, http://dx.doi.org/10.1016/J.INFFUS.2022.03.009.

[6] E. Siedlecka, T.F. Denson, Experimental methods for inducing basic emotions: A qualitative review, Emot. Rev. 11 (2019) 87–97, http://dx.doi.org/10.1177/1754073917749016, URL http://journals.sagepub.com/doi/10.1177/1754073917749016.

[7] B. Kurdi, S. Lozano, M.R. Banaji, Introducing the open affective standardized image set (OASIS), Behav. Res. Methods 49 (2017) 457–470, http://dx.doi.org/10.3758/S13428-016-0715-3/TABLES/2, URL https://link.springer.com/article/10.3758/s13428-016-0715-3.

[8] M. Gnacek, I. Mavridou, J. Broulidakis, C. Nduka, E. Balaguer-Ballester, T. Kostoulas, E. Seiss, AVDOS-affective video database online study video database for affective research emotionally validated through an online survey, in: 2022 10th International Conference on Affective Computing and Intelligent Interaction ACII 2022, Institute of Electrical and Electronics Engineers Inc., 2022, http://dx.doi.org/10.1109/ACII55700.2022.9953891.

[9] W. Yang, K. Makita, T. Nakao, N. Kanayama, M.G. Machizawa, T. Sasaoka, A. Sugata, R. Kobayashi, R. Hiramoto, S. Yamawaki, M. Iwanaga, M. Miyatani, Affective auditory stimulus database: An expanded version of the International Affective Digitized Sounds (IADS-E), Behav. Res. Methods 50 (2018) 1415–1429, http://dx.doi.org/10.3758/S13428-018-1027-6/TABLES/8, URL https://link.springer.com/article/10.3758/s13428-018-1027-6.

[10] V. Shuman, D. Sander, K.R. Scherer, Levels of valence, Front. Psychol. 4 (2013) 261, http://dx.doi.org/10.3389/fpsyg.2013.00261, URL www.frontiersin.org.

[11] J. Storbeck, G.L. Clore, Affective arousal as information: How affective arousal influences judgments, learning, and memory, Soc. Pers. Psychol. Compass 2 (2008) 1824–1843, http://dx.doi.org/10.1111/J.1751-9004.2008.00138.X, https://onlinelibrary.wiley.com/doi/full/10.1111/j.1751-9004.2008.00138.x https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-9004.2008.00138.x https://onlinelibrary.wiley.com/doi/10.1111/j.1751-9004.2008.00138.x.

[12] S. Bouchard, J. St-Jacques, G. Robillard, P. Renaud, Anxiety increases the feeling of presence in virtual reality, Presence: Teleoperators Virtual Environ. 17 (2008) 376–391, http://dx.doi.org/10.1162/PRES.17.4.376, URL https://dl.acm.org/doi/10.1162/pres.17.4.376.

[13] J.W. Elsey, K. van Andel, R.B. Kater, I.M. Reints, M. Spiering, The impact of virtual reality versus 2D pornography on sexual arousal and presence, Comput. Hum. Behav. 97 (2019) 35–43, http://dx.doi.org/10.1016/J.CHB.2019.02.031.

[14] D. Krupić, B. Žuro, P.J. Corr, Anxiety and threat magnification in subjective and physiological responses of fear of heights induced by virtual reality, Pers. Individ. Differ. 169 (2021) 109720, http://dx.doi.org/10.1016/J.PAID.2019.109720.

[15] I. Mavridou, E. Balaguer-Ballester, C. Nduka, E. Seiss, A reliable and robust online validation method for creating a novel 3D affective virtual environment and event library (AVEL), PLoS One 18 (2023) e0278065, http://dx.doi.org/10.1371/JOURNAL.PONE.0278065, URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0278065.

[16] P.D. Bernardo, A. Bains, S. Westwood, D.C. Mograbi, Mood induction using virtual reality: a systematic review of recent findings, J. Technol. Behav. Sci. 6 (2021) 3–24, http://dx.doi.org/10.1007/S41347-020-00152-9/TABLES/1, URL https://link.springer.com/article/10.1007/s41347-020-00152-9.

[17] M. Slater, Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments, Phil. Trans. R. Soc. B 364 (2009) 3549, http://dx.doi.org/10.1098/RSTB.2009.0138, URL /pmc/articles/PMC2781884//pmc/articles/PMC2781884/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2781884/.

[18] D. Liao, W. Zhang, G. Liang, Y. Li, J. Xie, L. Zhu, X. Xu, L. Shu, Arousal evaluation of VR affective scenes based on HR and SAM, in: IEEE MTT-S 2019 International Microwave Biomedical Conference, IMBioC 2019 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2019, http://dx.doi.org/10.1109/IMBIOC.2019.8777844.

[19] A.M. Brouwer, E. van Dam, J.B. van Erp, D.P. Spangler, J.R. Brooks, Improving real-life estimates of emotion based on heart rate: A perspective on taking metabolic heart rate into account, Front. Hum. Neurosci. 12 (2018) http://dx.doi.org/10.3389/fnhum.2018.00284.

[20] S. Jerritta, M. Murugappan, R. Nagarajan, K. Wan, Physiological signals based human emotion recognition: A review, in: Proceedings - 2011 IEEE 7th International Colloquium on Signal Processing and Its Applications, CSPA 2011, 2011, pp. 410–415, http://dx.doi.org/10.1109/CSPA.2011.5759912.

[21] B. J.E., A. S.V., K. A.H., N. D.J., Behavioral and cardiovascular effects of 7.5% CO2 in human volunteers, Depress. Anxiety 21 (2005) 18–25, http://dx.doi.org/10.1002/DA.20048, URL https://pubmed.ncbi.nlm.nih.gov/15782425/.

[22] S.Z. Poma, S. Milleri, L. Squassante, G. Nucci, M. Bani, G.I. Perini, E. Merlo-Pich, Characterization of a 7% Carbon Dioxide (CO2) Inhalation Paradigm to Evoke Anxiety Symptoms in Healthy Subjects:, Vol. 19, Sage PublicationsSage CA: Thousand Oaks, CA, 2016, pp. 494–503, http://dx.doi.org/10.1177/0269881105056533, URL https://journals.sagepub.com/doi/10.1177/0269881105056533.

[23] M. Garner, A. Attwood, D.S. Baldwin, A. James, M.R. Munafò, Inhalation of 7.5% carbon dioxide increases threat processing in humans, Neuropsychopharmacol. 36 (2011) 1557–1562, http://dx.doi.org/10.1038/npp.2011.15, 2011 36:8. URL https://www.nature.com/articles/npp201115.

[24] M. Gnacek, J. Broulidakis, I. Mavridou, M. Fatoorechi, E. Seiss, T. Kostoulas, E. Balaguer-Ballester, I. Kiprijanovska, C. Rosten, C. Nduka, emteqPRO—Fully integrated biometric sensing array for non-invasive biomedical research in virtual reality, Front. Virtual Real. 3 (2022) 3, http://dx.doi.org/10.3389/FRVIR.2022.781218.

[25] A.S. Cowen, D. Keltner, Self-report captures 27 distinct categories of emotion bridged by continuous gradients, Proc. Natl. Acad. Sci. 114 (2017) E7900–E7909, http://dx.doi.org/10.1073/PNAS.1702247114, https://www.pnas.org/content/114/38/E7900 https://www.pnas.org/content/114/38/E7900.abstract.

[26] M. Ménard, P. Richard, H. Hamdi, B. Daucé, T. Yamaguchi, Emotion recognition based on heart rate and skin conductance, in: Proceedings of the 2nd International Conference on Physiological Computing Systems - PhyCS, 2015, http://dx.doi.org/10.5220/0005241100260032.

[27] U. Dimberg, Facial electromyography and emotional reactions, Psychophysiology 27 (1990) 481–494, http://dx.doi.org/10.1111/J.1469-8986.1990.TB01962.X, URL https://pubmed.ncbi.nlm.nih.gov/2274612/.

[28] P. Ekman, W.V. Freisen, S. Ancoli, Facial signs of emotional experience, J. Pers. Soc. Psychol. 39 (1980) 1125–1134, http://dx.doi.org/10.1037/H0077722, URL /record/1981-25797-001.

[29] A.S. Cowen, H.A. Elfenbein, P. Laukka, D. Keltner, Mapping 24 emotions conveyed by brief human vocalization, Am. Psychol. 74 (2019) 698–712, http://dx.doi.org/10.1037/AMP0000399.

[30] I.-O. Stathopoulou, G.A. Tsihrintzis, Emotion recognition from body movements and gestures, Smart Innov. Syst. Technol. 11 SIST (2011) 295–303, http://dx.doi.org/10.1007/978-3-642-22158-3_29, URL https://link.springer.com/chapter/10.1007/978-3-642-22158-3_29.

[31] R.A. Calvo, S. D'Mello, Affect detection: An interdisciplinary review of models, methods, and their applications, IEEE Trans. Affect. Comput. 1 (2010) 18–37, http://dx.doi.org/10.1109/T-AFFC.2010.1.

[32] R.W. Picard, E. Vyzas, J. Healey, Toward machine emotional intelligence: Analysis of affective physiological state, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 1175–1191, http://dx.doi.org/10.1109/34.954607.

[33] J. Zhang, Z. Yin, P. Chen, S. Nichele, Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review, Inf. Fusion 59 (2020) 103–126, http://dx.doi.org/10.1016/J.INFFUS.2020.01.011.

[34] J.A. Domínguez-Jiménez, K.C. Campo-Landines, J.C. Martínez-Santos, E.J. Delahoz, S.H. Contreras-Ortiz, A machine learning model for emotion recognition from physiological signals, Biomed. Signal Process. Control. 55 (2020) 101646, http://dx.doi.org/10.1016/J.BSPC.2019.101646.

[35] S. D'Mello, J. Kory, Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies, in: Proceedings of the 14th ACM International Conference on Multimodal Interaction, 2012, pp. 31–38, http://dx.doi.org/10.1145/2388676.2388686.

[36] S.K. D'Mello, J. Kory, A review and meta-analysis of multimodal affect detection systems, ACM Comput. Surv. 47 (2015) http://dx.doi.org/10.1145/2682899, URL https://dl.acm.org/doi/abs/10.1145/2682899.

[37] M. Balsamo, L. Carlucci, C. Padulo, B. Perfetti, B. Fairfield, A bottom-up validation of the IAPS, GAPED, and NAPS affective picture databases: Differential effects on behavioral performance, Front. Psychol. 11 (2020) 2187, http://dx.doi.org/10.3389/fpsyg.2020.02187, URL https://www.frontiersin.org/article/10.3389/fpsyg.2020.02187/full.

[38] K.T.A. Baraly, L. Muyingo, C. Beaudoin, S. Karami, M. Langevin, P.S.R. Davidson, Database of emotional videos from Ottawa (DEVO), Collabra: Psychol. 6 (2020) http://dx.doi.org/10.1525/collabra.180, URL https://doi.org/10.1525/collabra.180,
.

[39] M. Slater, D.P. Pertaub, A. Steed, Public speaking in virtual reality: Facing an audience of avatars, IEEE Comput. Graph. Appl. 19 (1999) 6–9, http://dx.doi.org/10.1109/38.749116.

[40] M. C., D. S., On the validity of the autobiographical emotional memory task for emotion induction, PLoS One 9 (2014) http://dx.doi.org/10.1371/JOURNAL.PONE.0095837, URL https://pubmed.ncbi.nlm.nih.gov/24776697/.

[41] D. Wasserman, S.M. Liao, Issues in the pharmacological induction of emotions, J. Appl. Philos. 25 (3) (2008) 178–192, http://dx.doi.org/10.1111/j.1468-5930.2008.00414.x.

[42] X. Zhang, H.W. Yu, L.F. Barrett, How does this make you feel? A comparison of four affect induction procedures, Front. Psychol. 5 (2014) http://dx.doi.org/10.3389/FPSYG.2014.00689, /pmc/articles/PMC4086046//pmc/articles/PMC4086046/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4086046/.

[43] A.P. Soares, A.P. Pinheiro, A. Costa, C.S. Frade, M. Comesaña, R. Pureza, Adaptation of the international affective picture system (IAPS) for European portuguese, Behav. Res. Methods 47 (2014) 1159–1177, http://dx.doi.org/10.3758/S13428-014-0535-2, 2014 47:4. URL https://link.springer.com/article/10.3758/s13428-014-0535-2.

[44] C. Lasaitis, R.L. Ribeiro, O.F.A. Bueno, Brazilian norms for the international affective picture system (IAPS): comparison of the affective ratings for new stimuli between Brazilian and north-American subjects, J. Bras. de Psiquiatr. 57 (2008) 270–275, http://dx.doi.org/10.1590/S0047-20852008000400008, URL http://www.scielo.br/j/jbpsiq/a/VgFWQmVdLgkJknmWTPnmrrR/?lang=en.

[45] R. Adolphs, How should neuroscience study emotions? by distinguishing emotion states, concepts, and experiences, Soc. Cogn. Affect. Neurosci. 12 (2017) 24–31, http://dx.doi.org/10.1093/SCAN/NSW153, URL https://academic.oup.com/scan/article/12/1/24/2624554.

[46] F.M. MacDonald, Respiratory acidosis, Arch. Intern. Med. 116 (1965) 689–698, http://dx.doi.org/10.1001/ARCHINTE.1965.03870050043008, URL https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/572126.

[47] M. Garner, A. Attwood, D.S. Baldwin, M.R. Munafò, Inhalation of 7.5% carbon dioxide increases alerting and orienting attention network function, Psychopharmacology 223 (2012) 67–73, http://dx.doi.org/10.1007/S00213-012-2690-4/FIGURES/3, URL https://link.springer.com/article/10.1007/s00213-012-2690-4.

[48] G. Savulich, F.H. Hezemans, S. van Ghesel Grothe, J. Dafflon, N. Schulten, A.B. Brühl, B.J. Sahakian, T.W. Robbins, Acute anxiety and autonomic arousal induced by CO2 inhalation impairs prefrontal executive functions in healthy humans, Transl. Psychiatry 9 (2019) 1–10, http://dx.doi.org/10.1038/s41398-019-0634-z, 2019 9:1. URL https://www.nature.com/articles/s41398-019-0634-z.

[49] M. Pappens, S.D. Peuter, D. Vansteenwegen, O.V. den Bergh, I.V. Diest, Psychophysiological responses to CO2 inhalation, Int. J. Psychophysiol. 84 (2012) 45–50, http://dx.doi.org/10.1016/J.IJPSYCHO.2012.01.008.

[50] F. Safety, I. Service, Carbon dioxide health hazard information sheet, 2020, URL https://www.fsis.usda.gov/sites/default/files/media_file/2020-08/Carbon-Dioxide.pdf.

[51] E. Rostrup, H.B. Larsson, P.B. Toft, K. Garde, C. Thomsen, P. Ring, L. Søndergaard, O. Henriksen, Functional MRI of CO2 induced increase in cerebral perfusion, NMR Biomed. 7 (1994) 29–34, http://dx.doi.org/10.1002/NBM.1940070106, URL https://pubmed.ncbi.nlm.nih.gov/8068522/.

[52] F.B. Tancredi, R.D. Hoge, Comparison of cerebral vascular reactivity measures obtained using breath holding and CO2 inhalation, J. Cereb. Blood Flow Metab. 33 (2013) 1066–1074.

[53] S.V. Argyropoulos, J.E. Bailey, S.D. Hood, A.H. Kendrick, A.S. Rich, G. Laszlo, J.R. Nash, S.L. Lightman, D.J. Nutt, Inhalation of 35% CO2 results in activation of the HPA axis in healthy volunteers, Psychoneuroendocrinology 27 (2002) 715–729, http://dx.doi.org/10.1016/S0306-4530(01)00075-0, URL https://www.sciencedirect.com/science/article/abs/pii/S0306453001000750.

[54] AirFit F30 CPAP full face mask - ResMed, URL https://www.resmed.com/en-us/sleep-apnea/cpap-parts-support/sleep-apnea-full-products-list/cpap-masks/airfit-f30/.

[55] VIVE Pro Eye Overview | VIVE Southeast Asia, URL https://www.vive.com/sea/product/vive-pro-eye/overview/.

[56] Eye tracking technology for VR - VIVE Pro Eye with Tobii - Tobii, URL https://www.tobii.com/products/integration/xr-headsets/device-integrations/htc-vive-pro-eye.

[57] MP150 SYSTEMS, 2016, URL www.biopac.com.

[58] M.E. Kret, E.E. Sjak-Shie, Preprocessing pupil size data: Guidelines and code, Behav. Res. Methods 51 (2019) 1336–1342, http://dx.doi.org/10.3758/S13428-018-1075-Y/FIGURES/4, URL https://link.springer.com/article/10.3758/s13428-018-1075-y.

[59] I. Jackson, S. Sirois, Infant cognition: going full factorial with pupil dilation, Dev. Sci. 12 (2009) 670–679, http://dx.doi.org/10.1111/J.1467-7687.2008.00805.X, URL https://pubmed.ncbi.nlm.nih.gov/19635092/.

[60] P.J. Bota, C. Wang, A.L.N. Fred, H. Plácido Da Silva, A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals, IEEE Access 7 (2019) 140990–141020, http://dx.doi.org/10.1109/ACCESS.2019.2944001.

[61] M. Gnacek, L. Quintero, I. Mavridou, E. Balaguer-Ballester, T. Kostoulas, C. Nduka, E. Seiss, AVDOS-VR: Affective video database with physiological signals and continuous ratings collected remotely in VR, Sci. Data 11 (2024) 1–18, http://dx.doi.org/10.1038/s41597-024-02953-6, 2024 11:1. URL https://www.nature.com/articles/s41597-024-02953-6.

[62] M. Gjoreski, V. Janko, G. Slapničar, M. Mlakar, N.R. cič, J. Bizjak, V. Drobnič, M. Marinko, N. Mlakar, M. Luštrek, M. Gams, Classical and deep learning methods for recognizing human activities and modes of transportation with smartphone sensors, Inf. Fusion 62 (2020) 47–62, http://dx.doi.org/10.1016/J.INFFUS.2020.04.004.

[63] P. Jain, M. Rathee, Anatomy, head and neck, orbicularis oris muscle, StatPearls (2023) URL https://www.ncbi.nlm.nih.gov/books/NBK545169/.

[64] J. Pinto, M.G. Bright, D.P. Bulte, P. Figueiredo, Cerebrovascular reactivity mapping without gas challenges: A methodological guide, Front. Physiol. 11 (2020) http://dx.doi.org/10.3389/FPHYS.2020.608475, /pmc/articles/PMC7848198//pmc/articles/PMC7848198/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7848198/.

[65] J. Yamaguchi, K. Kinoshita, T. Hosokawa, S. Ihara, "The eyes are the windows of the soul": Portable automated pupillometry to monitor autonomic nervous activity in CO2narcosis: A case report, Med. (United States) 102 (2023) E33768, http://dx.doi.org/10.1097/MD.0000000000033768, URL https://journals.lww.com/md-journal/fulltext/2023/05120/_the_eyes_are_the_windows_of_the_soul__portable.29.aspx.

[66] W. Bourne, Lxxxix, 144; 1913, xCi, 126. (11) Roos, Zeitsch1r physiol. Chem., 1899, xxviii, 40; FONIO, Mitt, ASIIER, Dtsch. Med. Wochenschr (1921) 407.

[67] W. Sato, T. Kochiyama, S. Yoshikawa, Physiological correlates of subjective emotional valence and arousal dynamics while viewing films, Biol. Psychol. 157 (2020) 107974, http://dx.doi.org/10.1016/J.BIOPSYCHO.2020.107974.

[68] G. Verma, V. Vinay, R.A. Rossi, S. Kumar, Robustness of fusion-based multimodal classifiers to cross-modal content dilutions, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, 2022.

[69] S. Stankoski, I. Kiprijanovska, I. Mavridou, C. Nduka, H. Gjoreski, M. Gjoreski, Breathing rate estimation from head-worn photoplethysmography sensor data using machine learning, Sensors 22 (2022) 2079, http://dx.doi.org/10.3390/S22062079, 2022, Vol. 22, Page 2079. https://www.mdpi.com/1424-8220/22/6/2079/htm https://www.mdpi.com/1424-8220/22/6/2079.

[70] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, Inf. Fusion 99 (2023) 101805, http://dx.doi.org/10.1016/j.inffus.2023.101805.

[71] P. Romashov, M. Gjoreski, K. Sokol, M.V. Martinez, M. Langheinrich, BayCon: Model-agnostic Bayesian counterfactual generator, in: Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI, 2022, pp. 740–746.

[72] S. Koelstra, C. Mühl, M. Soleymani, J.S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, DEAP: A database for emotion analysis; using physiological signals, IEEE Trans. Affect. Comput. 3 (2012) 18–31, http://dx.doi.org/10.1109/T-AFFC.2011.15.

[73] K. Sharma, C. Castellini, E.L. van den Broek, A. Albu-Schaeffer, F. Schwenker, A dataset of continuous affect annotations and physiological signals for emotion analysis, Sci. Data 6 (2019) 1–13, http://dx.doi.org/10.1038/s41597-019-0209-0, 2019 6:1. URL https://www.nature.com/articles/s41597-019-0209-0.

[74] P. Bota, J. Brito, A. Fred, P. Cesar, H. Silva, A real-world dataset of group emotion experiences based on physiological data, Sci. Data 11 (2024) 1–17, http://dx.doi.org/10.1038/s41597-023-02905-6, 2024 11:1. URL https://www.nature.com/articles/s41597-023-02905-6.