Computers in Biology and Medicine 193 (2025) 110310

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Advancing emotion recognition with Virtual Reality: A multimodal approach using physiological signals and machine learning

Edoardo Maria Polo [a], Francesco Iacomi [a], Alberto Valdes Rey [a], Davide Ferraris [b], Alessia Paglialonga [c,*], Riccardo Barbieri [a]

[a] Politecnico di Milano, Piazza Leonardo da Vinci, 32, Milan, 20133, Italy
[b] Lega Italiana Per La Lotta Contro I Tumori – LILT, Via Giacomo Venezian, 1, Milan, 20133, Italy
[c] Cnr-Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni (CNR-IEIIT), Piazza Leonardo da Vinci, 32, Milan, 20133, Italy

ARTICLE INFO

Keywords:
Virtual Reality
Emotion recognition
Physiological signals
Machine learning
Explainable AI

ABSTRACT

Introduction : Emotion recognition systems have traditionally relied on basic visual elicitation. Virtual reality (VR) offers an immersive alternative that better resembles real-world emotional experiences.
Objective: To develop and evaluate custom-built VR scenarios designed to evoke sadness, relaxation, happiness, and fear, and to utilize physiological signals together with machine learning techniques for accurate prediction and classification of emotional states.
Methods: Physiological signals (electrocardiogram, blood volume pulse, galvanic skin response, and respiration) were acquired from 36 participants during VR experiences. Machine learning models, including Logistic Regression with Square Method feature selection, were applied in a subject-independent approach in order to discern the four emotional states.
Results: Features extracted by physiological signal analysis highlighted significant differences among emotional states. The machine learning models achieved high accuracies of 80%, 85%, and 70% for arousal, valence, and 4-class emotion classification, respectively. Explainable AI techniques further provided insights into the decision-making processes and the relevance of specific physiological features, with galvanic skin response peaks emerging as the most significant feature for both valence and arousal dimensions.
Conclusion: The proposed study demonstrates efficacy of VR in eliciting genuine emotions and the potential of using physiological signals for emotion recognition, with important implications for affective computing and psychological research. The non-invasive approach, robust subject-independent generalizability, and compatibility with wearable technology position this methodology favorably for practical applications in mental health contexts and user experience evaluation.

1. Introduction

1.1. The nature and classification of emotions

Emotions are complex phenomena that play a crucial role in human survival, adaptation, and social interaction. They involve intricate networks of neural and hormonal interactions, generating cognitive processes that influence decision-making and behavior [1]. Scherer's Component Process Model [2] elucidates the interplay between cognitive appraisal, physiological responses in the central and peripheral autonomic nervous systems (ANS), action tendencies, expressions, and feelings, highlighting the dynamic nature of emotional experiences. Understanding the neural, hormonal, and cognitive underpinnings of emotions is essential for explaining their influence on decision-making, perception, and interpersonal relationships.

Two primary approaches to emotion classification have emerged: the theory of basic emotions [3] and the dimensional models [4]. The former proposes a limited set of distinct emotional states, whereas the latter conceptualizes emotions as a combination of independent dimensions (i.e., valence, which refers to pleasantness, and arousal, which refers to excitement).

1.2. Physiological correlates and detection of emotions

The autonomic nervous system (ANS) regulates involuntary bodily functions and coordinates physiological responses to emotional stimuli,

* Corresponding author.
E-mail address: alessia.paglialonga@cnr.it (A. Paglialonga).

https://doi.org/10.1016/j.compbiomed.2025.110310
Received 26 November 2024; Received in revised form 25 April 2025; Accepted 29 April 2025
Available online 26 May 2025

with distinct emotions associated with specific activation patterns [5]. Emotion detection has advanced through the use of peripheral signals and electroencephalography (EEG) data, revealing correlations between brain activity patterns and emotional states, such as the association of high valence emotions with increased frontal alpha and right parietal beta power [6], and the link between parietal–occipital alpha power and arousal levels [7]. Heart Rate Variability (HRV) has been noted as an indicator of emotional states [8], showing distinct patterns of ANS activity in response to positive and negative stimuli [9]. Galvanic Skin Response (GSR) is particularly sensitive to high-arousal emotions, such as fear and disgust [10]. Respiratory patterns offer nuanced information, with high-valence emotions characterized by steady breathing [11]. The study of emotions and their physiological correlates has significant applications in human–computer interaction (HCI), human–robot interaction (HRI) [12], and clinical settings, including monitoring and treatment of depression, anxiety, stress, chronic anger, and mood disorders. [13]. Physiological signals, mainly associated with ANS, can be easily acquired through wearable devices, providing an objective approach to emotion detection [14]. Unlike self-report methods, which are subject to psychological biases (such as social desirability and recall bias) and practical limitations (including interruption of tasks for assessment and the inability to continuously monitor emotions), these physiological measures offer continuous, real-time, and objective monitoring of emotional states. By integrating data from multiple physiological systems and leveraging advances in machine learning and artificial intelligence, researchers can develop sophisticated emotion recognition systems with far-reaching implications in personalized healthcare, mental health interventions, and human–machine interaction.

1.3. Evolution of emotion elicitation: From flat screen to virtual reality

The evolution of emotion elicitation techniques has seen a significant transformation over the years, moving from actor portrayals [15] to standardized stimuli such as the International Affective Picture System (IAPS) [16], a large database of emotionally-evocative color photographs standardized on valence and arousal dimensions, ranging from pleasant to unpleasant and calm to excited scenes, and International Affective Digitized Sounds (IADS) [17], a standardized collection of acoustic stimuli designed to elicit emotional responses across the valence-arousal space, including naturalistic sounds, human vocalizations, and environmental noises. Both IAPS and IADS were developed by the Center for the Study of Emotion and Attention at the University of Florida and have been extensively validated across multiple cultures and age groups, providing researchers with calibrated tools for emotion induction in laboratory settings.[1] While these standardized stimuli have been widely used in emotion research, they are increasingly viewed as dated in the context of our modern, multisensory world. More recent datasets like DEAP [18], CASE [19], and MAHNOB-HCI [20] have attempted to incorporate more complex stimuli and physiological measurements, striving to balance experimental control with ecological validity.

In an era characterized by immersive digital experiences and constant multisensory stimulation, Virtual Reality (VR) has emerged as a compelling solution for emotion elicitation, offering significant advantages over traditional passive methods [21]. The immersive and interactive nature of VR enhances ecological validity and potentially evokes more genuine emotional responses, closely mimicking real-world experiences. The concept of "presence" in VR environments correlates significantly with the intensity of emotions experienced [22], presenting unique opportunities for more nuanced and realistic emotion recognition studies.

[1] Researchers interested in using these datasets can contact: CSEA Media Core Resources; email: media@cseamedia.org.

Research has shown that VR exposure therapy can be as effective as traditional in person therapy in treating a wide range of phobias, such as fear of heights, spiders, and flying [23]. The therapeutic applications of VR extend beyond simple phobias to include more complex psychological conditions, such as post-traumatic stress disorder among military veterans and body image disorders [24]. In educational contexts, VR-based interventions have shown promise in rehabilitating children with dyslexia [25]. Recent research has explored VR's potential in emotion elicitation and recognition, with studies achieving satisfactory accuracy in classifying emotions along valence and arousal dimensions. For instance, [26] involved 12 participants and four emotions, utilizing pre-existing VR scenes. Another study, [27] achieved an 84% accuracy rate in distinguishing between high and low cognitive control levels using 26 participants and pre-selected VR videos from various sources. The AVDOS-VR dataset brought about a notable advancement [28], which features 30 emotionally labeled videos and two physiological signals collected remotely from 37 subjects, aiming to bridge the gap between controlled laboratory settings and real-world environments. More recent studies have further explored emotion recognition in VR environments. For example, [29] utilized EEG measurements with VR-projected videos to classify three emotions (including neutral) with 77.75% accuracy and four emotions (including neutral) with 62.67% accuracy. However, this approach relied on pre-existing video content rather than custom-designed VR environments specifically tailored for emotion elicitation. Additionally, while EEG provides valuable neural data, complexity and potential discomfort limit practical applications in real-world settings where wearable devices would be preferable. Another recent approach [30] achieved high performance using facial expression recognition with 2D images for interactive VR applications and neural network modeling. Nevertheless, facial monitoring presents limitations regarding privacy concerns and may not always accurately reflect internal emotional states, as expressions can be consciously masked — this is particularly relevant in clinical populations where emotional expression and experience may be dissociated. Despite these advancements, current VR-based emotion datasets often rely on pre-existing content and may have limited diversity in recorded physiological signals. As the field evolves, there is a growing need for more comprehensive datasets that incorporate custom-designed VR environments and a wider range of physiological signals to enhance the sense of presence and ecological validity. These limitations underscore the need for emotion recognition approaches that combine custom-designed VR environments with non-invasive physiological measurements to enhance ecological validity while maintaining practical applicability in real-world contexts.

1.4. Development of a novel and comprehensive VR-based emotion recognition protocol

Our study addresses these limitations by developing a novel VR protocol specifically designed for emotion recognition. We created novel immersive scenarios to elicit four distinct emotions, acquiring four physiological signals from 36 participants - a larger sample size compared to many existing studies. Unlike previous research that may achieve high classification accuracies without providing detailed information on the features used, our study offers comprehensive feature analysis and explainability processing of signals. Moreover, our protocol was developed in collaboration with a psychologist, ensuring its psychological validity. This multidisciplinary approach, combined with the use of VR technology, allows for precise control over stimuli while maintaining high ecological validity. Importantly, our study utilizes 5 min intervals for physiological recordings, surpassing the minimum 30-second threshold required for robust HRV analysis [31].

Preliminary findings of this research were presented in [32].

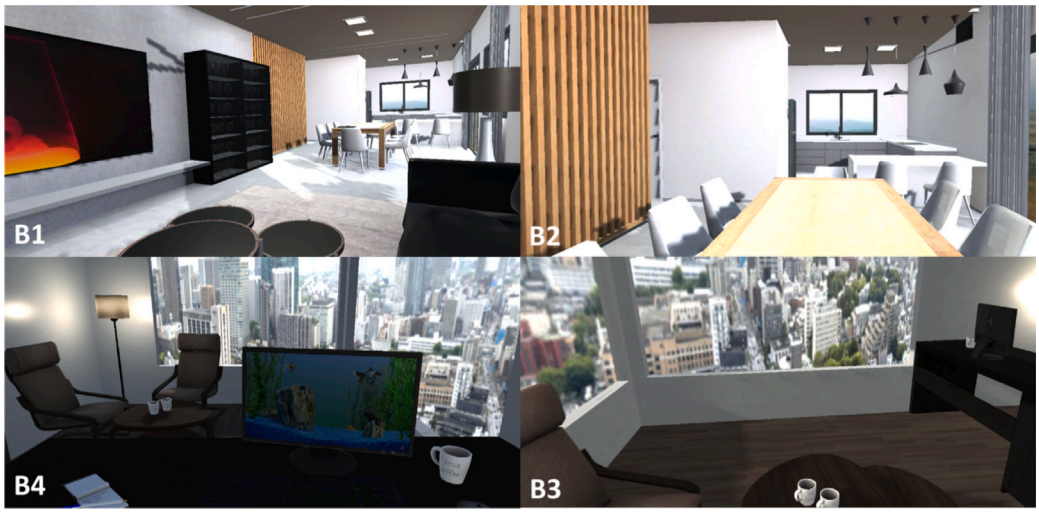

Fig. 1. The 4 neutral rooms developed in Unity to separate each emotional room from the others.

2. Methods

2.1. Study design and data

The study cohort comprised 36 Italian university students (33.3% female) aged between 22 and 30 years (24.29 ± 1.99 years). The experimental protocol was conducted at the SpinLab facility of Politecnico di Milano, following written informed consent from all participants, and was approved by the Politecnico di Milano Research Ethics Committee (Approval No. 29/2021, date: 14/07/2021). During the VR-based emotional elicitation procedure, physiological signals including electrocardiogram (ECG), blood volume pulse (BVP), GSR, and respiratory patterns (RESP) were continuously monitored. All biosignals were acquired using the Procomp Infinity system (model SA7500), with sampling rates of 2048 Hz for ECG and BVP, and 256 Hz for GSR and RESP. Participants were instructed to abstain from caffeine consumption and smoking for a minimum of two hours prior to the experimental session. Exclusion criteria were applied to individuals reporting pre-existing medical conditions or psychiatric disorders to maintain homogeneity within the study population.

2.2. The VR emotional experience

A fully immersive VR protocol was designed to evoke sadness (S), relaxation (R), happiness (H), and fear (F). We focused on these four main emotions for several reasons: to ensure distinct separation between emotional states along valence and arousal dimensions, to maintain reasonable experiment duration to prevent participant fatigue, and to focus on emotions with established physiological signatures in the literature. These emotions were specifically selected to cover the full range of valence and arousal combinations described by Russell's circumplex model of affect. The protocol followed an increasing arousal strategy, alternating emotional states with neutral environments, to prevent higher-arousal emotions from influencing the physiological responses to lower-arousal states [33,34].

2.2.1. VR environment design

The VR experience was developed using Unity (version 2020.3.22f1) for compatibility with future devices and software, allowing complete control over the experience. Key features include:

- Audio–visual cues were selected based on an extensive study of affective states cues from previous investigations [35]. All scenes were inspired by the validated IAPS and IADS datasets. Following IAPS guidelines [36] and their standardized valence-arousal framework, we designed VR environments to match characteristic stimuli patterns for each emotion: loss scenarios for S (low arousal/negative valence), threatening elements similar to accident imagery for F (high arousal/negative valence), natural landscapes for R (low arousal/positive valence), and food presentations with social interactions for H (high arousal/positive valence). IADS sounds were matched to ensure congruence with visual stimuli, with auditory elements falling within the same circumplex quadrant as their associated visual components to reinforce the targeted emotional states.
- Participants could rotate their heads within 180° to enhance immersion while minimizing motion artifacts.
- The user was represented by a seated avatar, enabling visual perception of their virtual body and hands.
- Direct interaction with virtual objects was restricted to isolate the effects of visual and auditory stimuli, reducing confounding variables.

To mitigate potential challenges associated with VR-based emotion research, we implemented several key strategies to ensure participant comfort and engagement. Our VR scenarios were carefully designed with smooth, controlled camera movements and minimalist, non-distracting environments to reduce cognitive load and prevent cybersickness. We conducted individual headset fittings and maintained real-time monitoring of participant comfort levels, allowing for immediate adjustments if needed.

2.2.2. Neutral environments

The protocol begins with a two-minute adaptation room to mitigate positive bias in valence and arousal provoked by VR. In this scene, users can interact with a dice on a desk. Rather than monochromatic baselines, neutral rooms (B1–B4, shown in Fig. 1) minimize stimuli while maintaining immersion. They reproduce real-world neutral scenarios with mixed stimuli to elicit neutral responses:

- Living room-like space (B1, B2) and office (B3, B4)
- Angular shapes, straight lines, unsaturated gray hues
- Slow-moving visual cues (e.g., images on monitors)
- Relaxing, slow-tempo music with intermittent bird sounds

2.2.3. Psychological foundation

The design of emotional scenes was grounded in a comprehensive psychological framework:

- Sadness: Designed to evoke isolation, loneliness, and social disconnection, drawing from cognitive theories of depression and the social nature of sadness [37,38].
- Relaxation: Conceptualized around natural settings, supported by research on nature's restorative effects and mindfulness-based stress reduction [39,40]. Incorporated attention restoration theory.

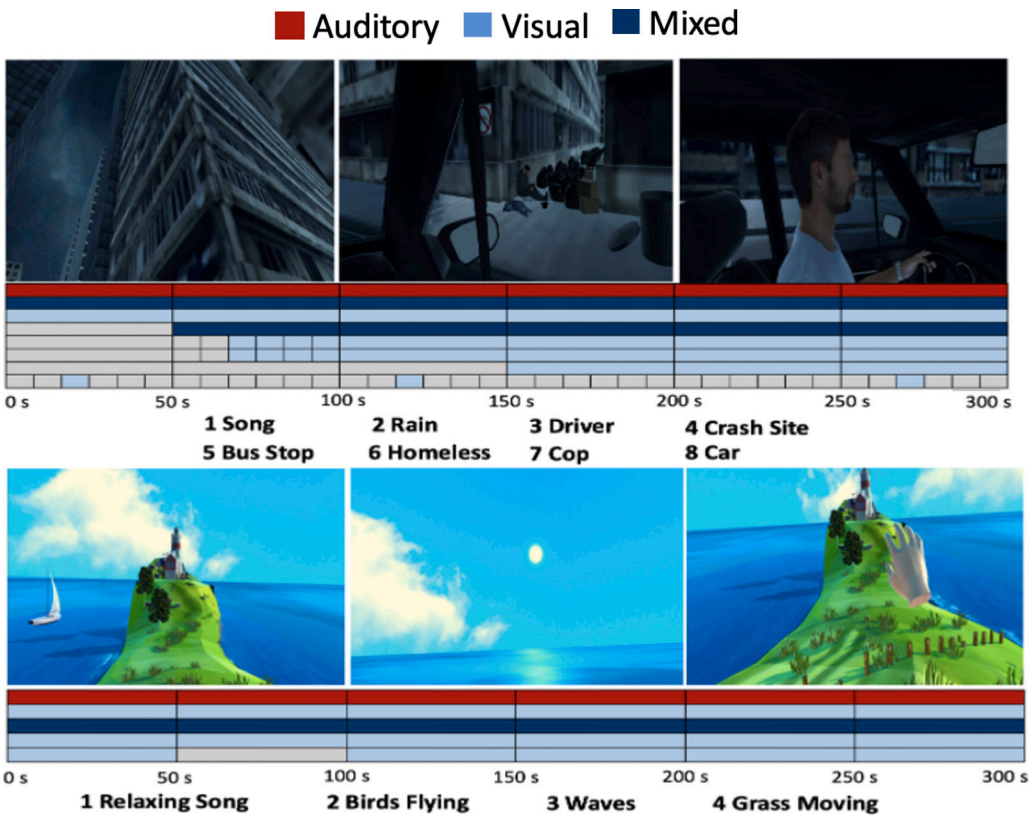

Fig. 2. Screenshots and stimulus timelines for low arousal VR emotional scenes (S and R). Three images are provided for each scene, accompanied by timelines categorizing video, audio, and mixed stimuli. In the timelines, each row corresponds to the stimulus listed directly below it, with the stimuli ordered from top to bottom (i.e., the first row represents stimulus 1, the second row represents stimulus 2, and so on).

- Happiness: Focused on social interaction, interpersonal connections, and shared positive experiences, grounded in positive psychology and theories of social well-being [41,42].
- Fear: Incorporated unexpected threats, death anxiety, phobias, based on evolutionary theories of fear and cognitive models of anxiety. Emphasized uncertainty and lack of control [43,44].

2.2.4. Emotional scene design

Each of the four emotional scenes was designed to have a duration of 5 min. The scenes are presented in the following order:

- Sadness: Set in a car during a rainy accident. Angular shapes, straight lines, purple–blue hue, low brightness. Rain, engine sounds, crying, sad music.
- Relaxation: Wild green island scene [45]. Rounded shapes, blue–green hues, high brightness. Waves, breeze sounds.
- Happiness: Kitchen party with food and drinks. Mixed shapes, complex environment, green–yellow hues, high saturation. Glass bottles, happy music, chatter.
- Fear: Old haunted house bathroom during thunderstorm [46]. Paranormal activities, angular shapes, red–yellow hues, dark, unsaturated. Loud, unsettling sounds.

Figs. 2 and 3 present screenshots and stimulus timelines for each VR emotional scene.

2.2.5. Subjective assessment

To assess the subjective effect of the protocol, a survey was designed using Google Forms and administered immediately after the VR experience. The survey presented screenshots of each virtual scene, followed by the Pick-A-Mood (PAM) model [47], where subjects were asked to associate the corresponding emotion. This assessment was conducted for both visual and auditory components of the experience.

2.3. Data processing and feature extraction

All data processing and feature extraction steps were performed using MATLAB R2022b. In the following subsections the processing and the feature extraction steps are explained for each signal. The complete list of features, along with their calculation methods and descriptions, can be found in Table A of the Supplementary Material.

2.3.1. ECG

2.4. ECG signal processing and feature extraction

The raw ECG signal underwent preprocessing with a Butterworth low-pass filter (125 Hz cutoff) and was downsampled to 250 Hz. We detected R-peaks using the Pan-Tompkins algorithm [48] and manually verified all annotations to ensure accuracy. From these R-peaks, we constructed the tachogram (RR interval time series) for further analysis.

2.4.1. Time domain and nonlinear features

From the tachogram, we extracted 13 features. The SLOPE parameter captured the heart rate trend through linear regression over the 5 min window. Basic statistical measures included the mean of RR intervals (AVNN, Average of the NN intervals), standard deviation of RR intervals (SDNN, Standard Deviation of the NN intervals), and variance of RR intervals (Var). The mean of the standard deviations of NN (Normal-to-Normal) intervals across 5 min segments was represented by SDNNINDX (Standard Deviation of the NN Index). For beat-to-beat variability, we calculated RMSSD (root mean square of successive differences), its logarithm (logRMSSD), and SDSD (standard deviation of successive differences). The nonlinear analysis included correlation dimension (CorrDim), Hurst exponent (H), and maximum Lyapunov exponent (LyapEx) to characterize signal complexity. Standard deviations

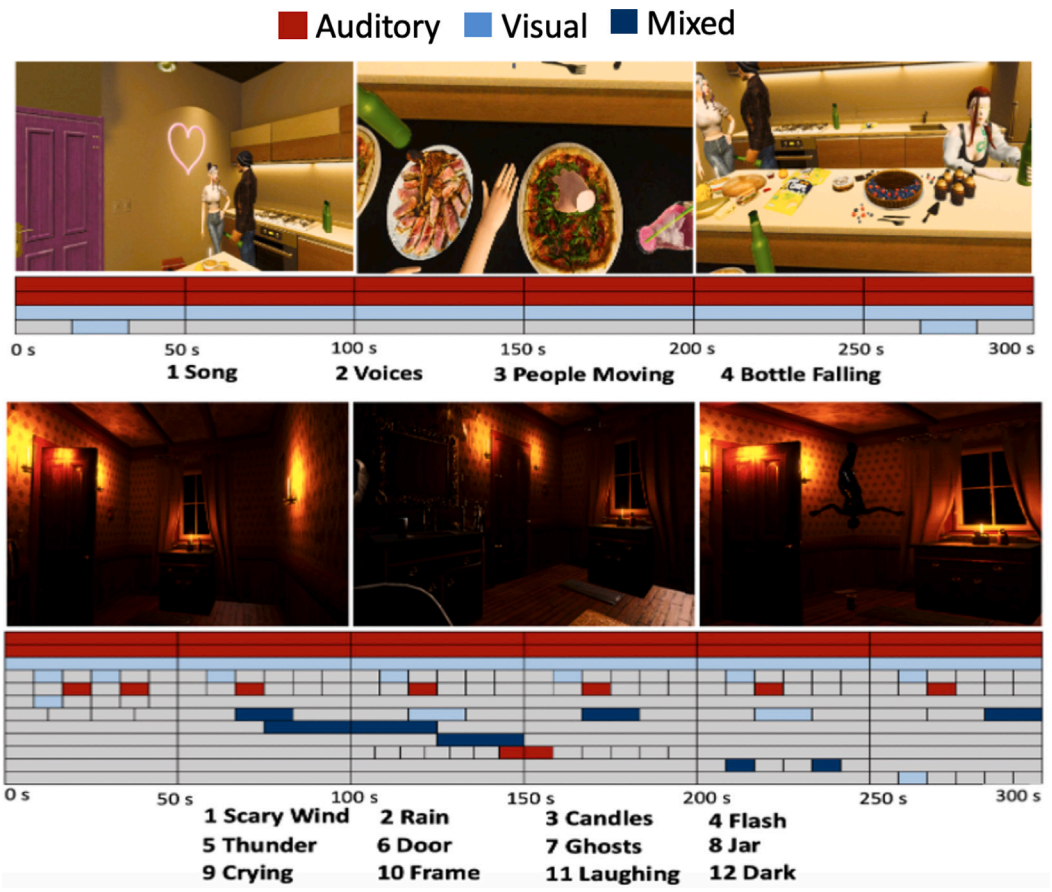

Fig. 3. Screenshots and stimulus timelines for high arousal VR emotional scenes (H and F). Three images are provided for each scene, accompanied by timelines categorizing video, audio, and mixed stimuli. In the timelines, each row corresponds to the stimulus listed directly below it, with the stimuli ordered from top to bottom (i.e., the first row represents stimulus 1, the second row represents stimulus 2, and so on).

from Poincaré plot calculated by applying Principal Component Analysis (PCA) to the RR interval scatter plot (RR(n) vs. RR(n+1)), where SD1 and SD2 represent the standard deviations from the first two principal components, representing short-term and long-term variability patterns.

2.4.2. Frequency domain features

We performed spectral analysis to examine frequency-specific components. We calculated the total power spectral density (PSD) (RR TOTPWR) and PSD in three specific bands: very low frequency (RR VLF, <0.04 Hz), low frequency (RR LF, 0.04–0.15 Hz), and high frequency (RR HF, 0.15–0.5 Hz). These bands reflect different aspects of autonomic regulation. We also derived the LF/HF ratio (RR LFtoHF) and normalized units (RR LFnu, RR HFnu) to assess sympathovagal balance. The RR spect slope was computed from the Lomb–Scargle periodogram to characterize the overall spectral distribution.

2.4.3. Point process framework

To handle the non-stationary nature of emotional responses, we implemented a Point Process framework [49]. This approach models heartbeats as point events in time, providing continuous estimates of heart rate dynamics [50]. The model yielded time-varying estimates of mean (μRR) and variance (σ^2) of RR intervals, along with spectral measures (RR tot, vlf, lf, hf). We also computed the instantaneous sympathovagal balance (RR lf/hf) and normalized spectral components (RR lfn, hfn), allowing us to track autonomic changes throughout the emotional stimuli. From the frequency and point process domains a total of 17 features were extracted.

2.4.4. ECG features: Physiological relevance for emotion recognition

The extracted ECG features provide valuable physiological insights into ANS dynamics during emotional processing. Time domain measures (AVNN, SDNN, RMSSD) primarily quantify parasympathetic activity, which typically decreases during negative emotional states [51]. Nonlinear indices (CorrDim, Hurst exponent, LyapEx) characterize the complexity reduction observed in cardiovascular dynamics during emotional stress [33]. Frequency domain parameters offer complementary information, with HF components reflecting parasympathetic regulation, LF representing mixed sympathetic–parasympathetic influences, and LF/HF ratio serving as an index of sympathovagal balance that increases during negative states [52]. The point process framework provides instantaneous assessment of autonomic regulation throughout emotional responses, capturing the dynamic temporal patterns that distinguish different affective states [34].

2.4.5. BVP

2.5. BVP signal processing and feature extraction

The BVP signal underwent preprocessing to remove high-frequency noise while preserving essential morphological features. A fourth-order zero-phase low-pass Butterworth filter with a 25 Hz cut-off frequency was applied, followed by downsampling to 250 Hz. Fiducial points, including systolic peaks, diastolic troughs, and pulse wave onsets, were extracted in relation to ECG R-peak annotations.

2.5.1. BVP waveform correction

To address temporal and amplitude anomalies in the BVP waveform, which often arose from irregularities in ECG R-peak annotations

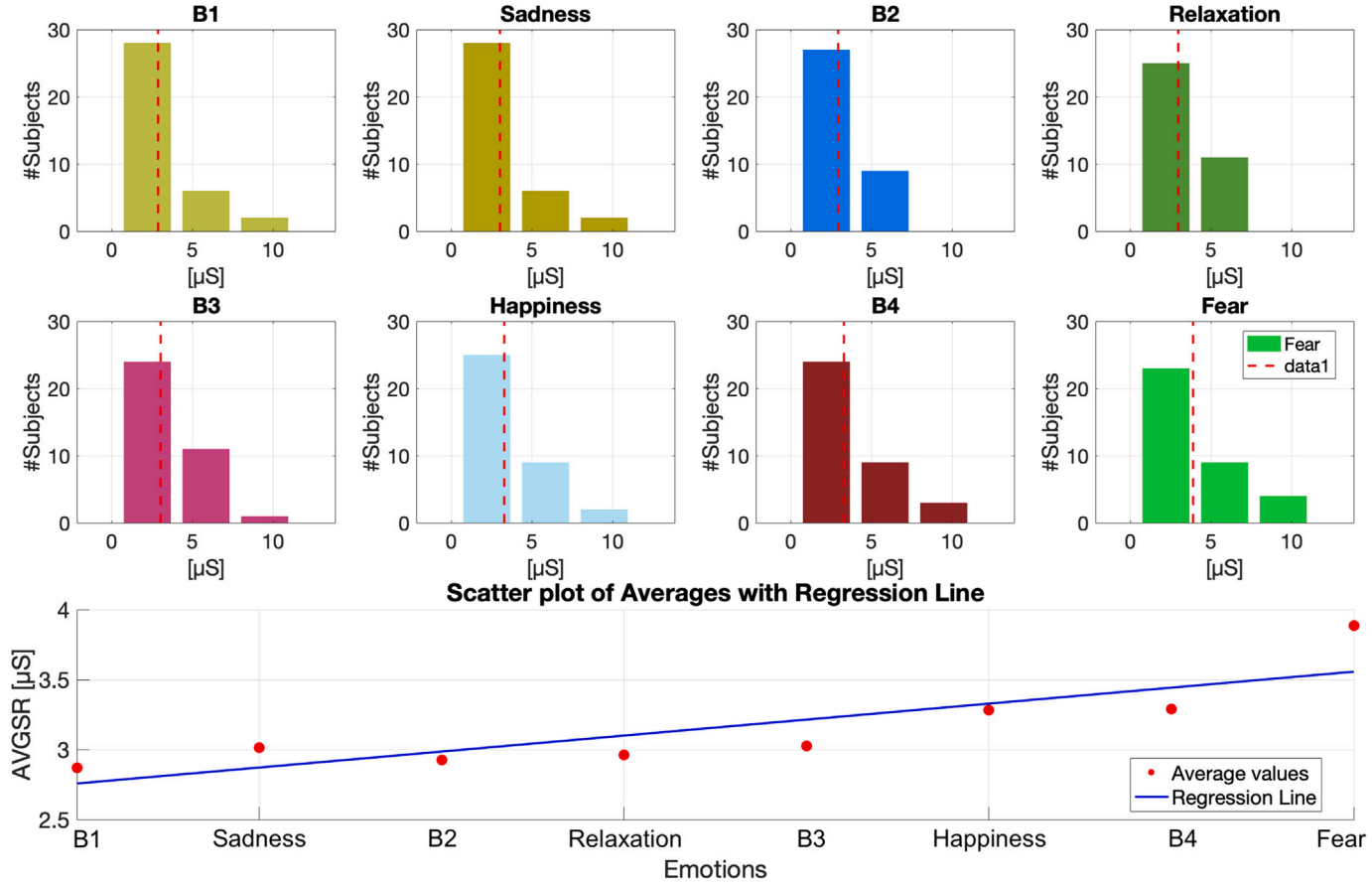

Fig. 4. (a) Histograms of eight distinct VR rooms, each displayed with three bins, showing the distribution of values for AVGSR. The histograms are plotted side by side, each indicating the frequency of data points within uniform value ranges across all emotional rooms. The red dashed vertical lines represent the mean values for each group. (b) Scatter plot of the mean values from each group, with the x-axis representing the VR rooms (1 to 8) and the y-axis depicting the median values. The blue regression line highlights the linear relationship between the VR rooms and their corresponding mean values of AVGSR.

or atypical morphological distortions, a correction algorithm was developed. This algorithm leveraged the physiological relationship between ECG and BVP signals, employing adaptive estimation techniques for BVP fiducial points and weighted averaging schemes to maintain physiological plausibility in the corrected signal.

2.5.2. BVP-derived features

Several features were extracted from the BVP signal and its relationship with the ECG within the emotion time windows. A total of 5 BVP features were computed.

The average volume pulse (*VP*) was computed as the average amplitude difference between each systolic and the corresponding diastolic value, serving as an index of the average blood volume pulse. The variance of the BVP time series (*VAR BVP*) was calculated to quantify BVP variability. The average pulse-to-pulse interval (*PulsePulse*) was obtained by averaging the onset-to-onset time series, expressed in milliseconds. The pulse arrival time (*PAT*) was computed as the average time difference between each onset value on the BVP signal and the corresponding R-peak on the ECG signal. Finally, the overall trend of the BVP signal was represented by the *Slope BVP* parameter, derived using first-degree polynomial fitting.

2.5.3. BVP features: Physiological relevance for emotion recognition

The BVP features provide critical information about peripheral vascular dynamics during emotional responses. VP reflects vasomotor activity and blood displacement volume, and represents, indeed, the volume of blood on the periphery and lower amplitudes are therefore linked to a greater peripheral blood pressure, which is associated with vasoconstriction and used in emotion recognition [53,54]. VAR BVP quantifies the stability of peripheral blood flow, which shows characteristic patterns of decreased variability during negative states [54]. PulsePulse measures cardiac chronotropic regulation similar to RR intervals but from a peripheral perspective, offering complementary information about autonomic control during emotional processing. PAT serves as an indirect measure of blood pressure changes and arterial stiffness, with shorter values observed during stress and fear due to sympathetic activation. The Slope BVP parameter captures the overall trend in peripheral blood volume over time, revealing sustained vasoconstriction or vasodilation patterns.

2.5.4. GSR

2.6. GSR signal processing and feature extraction

The GSR signal was preprocessed using a 4th order zero-phase low-pass Butterworth filter with a cut-off frequency of 2 Hz and downsampled to 5 Hz. To separate phasic and tonic components, a median filter was applied, replacing each sample amplitude with the median amplitude of the surrounding samples in a time window of +/−4 s centered on the current sample [55]. The resulting median signal was subtracted to obtain the phasic signal, which was used to extract GSR peaks linked to eccrine glands' spikes. GSR peaks were computed as local maxima on the filtered signal between each onset (amplitude > 0.01 μS) and offset (0 μS < amplitude) occurrences on the phasic signal.

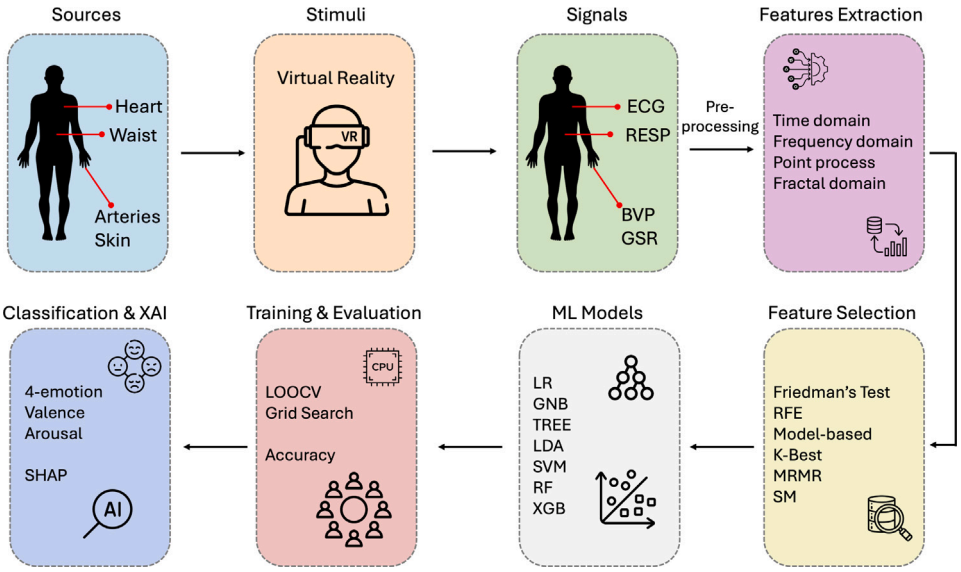

Fig. 5. General pipeline of the study.

2.6.1. GSR-derived features

Given the slower time scale of the GSR signal and the need to consider relatively short time windows for analysis, only time-domain features were extracted.

The extracted features include the number of GSR peaks identified within the signal (GSR n peaks), the average amplitude and standard deviation of the detected GSR peaks (GSR Amp peaks and GSR SD Amp peaks, respectively), and the average rise and recovery times of the GSR peaks (AVGSR rise time and AVGSR recovery time, respectively). Statistical measures such as the average and standard deviation of the GSR signal (AVGSR and SDGSR, respectively) were also computed. Additional features were derived from the filtered and transformed GSR signal, including the maximum signed amplitude between two consecutive maxima (MAXGSR sign amp), the average, standard deviation, and maximum of the first derivative of the GSR signal (AVGSR der, SDGSR der, and MAXGSR der, respectively), and the envelope of the phasic component (AVenv). The slope of the band-pass filtered GSR signal (Slope GSR) and the mean values of the absolute first and second derivatives of the low-pass filtered GSR (Avg abs1 and Avg abs2), along with their normalized counterparts (Avg abs1 norm and Avg abs2 norm), were also calculated to capture both tonic and phasic components of the GSR signal.

2.6.2. GSR features: Physiological relevance for emotion recognition

The extracted GSR features capture critical aspects of sympathetic nervous system activation during emotional processing. Tonic and phasic components of the signal provide complementary information about emotional arousal levels [15,56]. The average conductance and its statistical derivatives reflect sustained sympathetic activity, with higher values consistently associated with increased emotional arousal regardless of valence [57]. Temporal characteristics of GSR peaks, including amplitude and rise time, demonstrate distinct patterns across different emotional states, particularly discriminating high-arousal emotions from low-arousal [58]. The recovery time of GSR responses further differentiate between emotions of similar arousal levels but different valence [59]. Collectively, these features enable robust classification of affective states by quantifying the intensity and temporal dynamics of sympathetic arousal during emotional experiences.

2.6.3. RESP

The respiratory signal was preprocessed using the Parks-McClellan algorithm, a zero-phase digital low-pass filtering [60], in order to isolate the relevant frequency components. The filter was designed with a cutoff frequency of 1 Hz. Following filtering, peaks and troughs were detected using a prominence-based thresholding technique. For peaks (local maxima), the minimum peak prominence was set to 0.3 times the signal's standard deviation. Troughs (local minima) were identified using the same approach by inverting the signal polarity, thus applying the same prominence threshold to the negative peaks. Two key features were extracted from the processed RESP signal: the respiratory frequency *fRESP*, and the mean breath amplitude *ampRESP*. The latter was computed as the mean difference between maxima and minima of consecutive breath cycles.

2.6.4. RESP features: Physiological relevance for emotion recognition

Respiratory features provide valuable indices of emotional arousal and valence through autonomic modulation of breathing patterns. fRESP serves as a sensitive marker of emotional activation, with increased rates observed during high-arousal states such as fear and anxiety and decreased during low-arousal [61]. ampRESP reflects the depth of respiration, which typically during negative emotional states, particularly those involving sympathetic activation [62].

2.7. Dataset description

The dataset consists of data collected from 36 subjects, each experiencing the four target emotions: S, R, H, and F. Consequently, the dataset contains 36 * 4 = 144 rows, with each row representing a specific subject-emotion combination. The columns in the dataset represent the features calculated from the preprocessed physiological signals (ECG, BVP, GSR, and RESP). The total number of columns is 55, which is the sum of all the features extracted from these signals: 30 features from ECG, 5 features from BVP, 17 features from GSR, and 2 features from RESP. Additionally, there is a label column indicating the corresponding emotion for each row, using the labels S, R, H, and F. It is important to note that each feature is calculated as the difference between the feature value for the corresponding emotion and the feature value for the preceding neutral scene. This approach isolates the emotional response from the baseline physiological state. The rationale behind this detrending method will be further explained in the following Section 2.8.

2.8. Statistical analysis

The experimental protocol presented emotional stimuli in a sequence of increasing arousal, which was considered in the statistical analysis and classification procedures. Fig. 4 focuses on the AVGSR feature, a GSR feature known to be associated with emotional arousal [63,64], to demonstrate the effectiveness of the experimental protocol in eliciting the intended emotional responses. By visualizing the increase in AVGSR as the protocol progresses from low-arousal to high-arousal emotions, we provide evidence supporting the protocol's design. Moreover, the AVGSR feature serves as a representative example to illustrate the detrending process applied to all 54 features. Detrending each feature with respect to the preceding neutral scene isolates the emotional response from baseline physiological states, controlling for individual variability and time-dependent effects. While Fig. 4 showcases the AVGSR feature due to its established link to arousal, the same detrending process and statistical analyses were applied to all features extracted from the ECG, BVP, and GSR signals, as described in the previous section. The normality of the data distributions was assessed using the Shapiro–Wilk test, and a non-parametric Friedman's test was employed for pairwise comparisons due to deviations from normality. Post-hoc analyses using a multiple comparison test with Bonferroni correction were conducted to identify specific phases with significant variations. This comprehensive statistical approach ensures robust identification of significant physiological changes across emotional states while accounting for the sequential nature of the experimental design and controlling for potential confounds.

2.9. Machine learning

Following the statistical analysis outlined in Section 2.8, we developed a machine learning pipeline to address three classification problems: (i) emotion recognition (4 classes); (ii) high/low arousal detection (binary classification), and (iii) high/low valence detection (binary classification). Fig. 5 presents a flowchart illustrating the protocol pipeline from the acquisition of physiological signals to classification. The same dataset was considered across all three classification problems, with only the target variable being modified. This variable comprised, respectively, the four emotions, low and high arousal (combining S with R and H with F), and low and high valence (combining S with F and R with H).

After the detrending process described in Section 2.8, the dataset underwent cleaning and splitting. Initially, we partitioned the dataset into training (80%, i.e., 29 subjects) and test sets (20%, i.e., 7 subjects) based on subject allocation, ensuring that each subject was exclusively assigned to either the training or test set, thus preventing any subject overlap. Subsequently, we applied a standard scaler to normalize each variable by removing the mean and scaling to unit variance. The scaler was fitted on the training set and then used to transform both the training and test sets.

2.9.1. Feature selection methods

We considered six distinct feature selection methods to identify the most relevant features on the training set. These methods are outlined below:

- *Friedman's Test*: This non-parametric statistical test assesses the significance of features across different target classes. We applied the Friedman's test to each variable, grouped by target, and eliminated features with p-values > 0.05. Subsequently, we removed highly correlated features (correlation > 0.8), retaining the most significant feature based on the test statistic Q.
- *Recursive Feature Elimination (RFE)*: This iterative feature selection method uses a Random Forest (RF) estimator to recursively remove the least important features. It starts with all features and progressively eliminates the least significant ones based on the estimator's feature importance rankings.
- *Ensemble Method*: This approach combines three model-based techniques: RF feature importance ranking, Sequential feature selection using Logistic Regression, Permutation importance method with Logistic Regression (LR). We retained only the variables selected by at least two of these methods, leveraging the strengths of multiple approaches.
- *K-Best Selection*: This method selects the k best features based on the ANOVA F-value for the provided sample. It assesses the relationship between features and the target variable, ranking features by their F-scores. We set $k = 40$ for our analysis.
- *Minimum Redundancy Maximum Relevance (MRMR)*: This algorithm selects features that are maximally relevant to the target variable while minimizing redundancy among selected features. It balances the trade-off between relevance and redundancy, aiming to create a diverse and informative feature set. We configured the algorithm to select 40 features.
- *Square Method (SM)*: This method aims at identifying the most significant features for optimal class separation in a 2D plane. It creates 2D boxes for each feature pair and class, calculates their overlap, and assigns importance based on minimal intersection. The method then weights features by their frequency in non-discarded pairs, eliminates highly correlated features, and selects those with above-average weights [65]. This approach yields simple, interpretable methods by focusing on the most discriminative feature combinations and provides a multi-faceted evaluation of feature importance.

2.9.2. Classifiers

For each set of selected variables, we trained seven machine learning models among the most frequently used for emotion recognition [66,67] including: LR, Gaussian Naive Bayes (GNB), Decision Tree Classifier (DT), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), RF and XGBoost (XGB).

Each model was trained using Leave-One-Out Cross-Validation (LOOCV), wherein one patient at a time was used to validate the hyperparameters. These hyperparameters were optimized via Grid Search choosing ranges to be broad enough to explore different configurations while carefully mitigating overfitting. This approach ensures robust model evaluation by testing the model's performance on unseen data. The metrics used to evaluate the models' performance are: accuracy, F1-score, recall, and precision. However, since each classification problem is perfectly balanced and all classes have equal importance, we selected accuracy as the primary metric for identifying the top-performing models. All the algorithms for the analyses were coded using Python 3.10.13.

2.9.3. Explainable artificial intelligence (XAI)

After conducting the three classification tasks, we applied Explainable AI (XAI) methods to the best-performing model with the optimal feature selection method. This step is crucial as it provides a new perspective on the classification process beyond mere model performance. It allows us to understand the decision-making process of the model and gain deeper insights into the role of each variable.

Specifically, we employed SHapley Additive exPlanations (SHAP) [68], a method based on cooperative game theory. SHAP assigns to each feature an importance value for the specific prediction purpose (i.e., arousal or valence). These values represent the impact of a feature on the model output, with positive values indicating a feature that increases the likelihood of a specific output class, and negative values indicating the opposite. The magnitude of the SHAP value can be interpreted as the feature's importance in the decision.

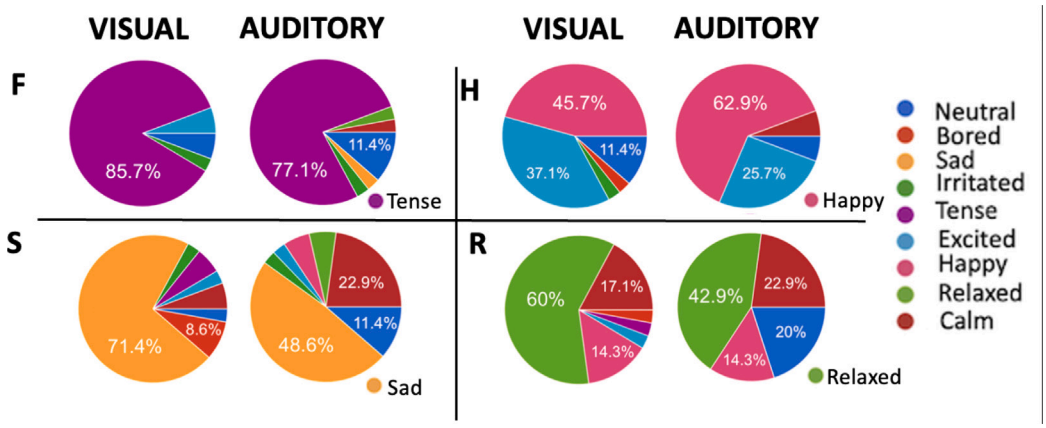

Fig. 6. Pie charts depicting the emotional labeling of the four VR emotional scenes, divided into visual (video) and auditory (audio) stimulation. The percentages represent the proportion of subjects who selected each emotion during the post-experiment PAM questionnaire.

3. Results

3.1. Subjects' labeling of VR rooms

The PAM behavioral responses provided by the participants revealed a strong alignment between the intended emotions during the protocol design phase and the subjective judgments of the stimuli. Fig. 6 presents pie charts illustrating the emotional labeling of auditory and visual stimuli for each scene, demonstrating strong agreement between the intended and perceived emotions. The fear-inducing scenario achieved the highest consensus among participants for both visual and auditory components, suggesting that fear-inducing stimuli may be more universally recognized or experienced across individuals. In general, with the exception of the happiness scene, there was greater agreement for visual stimuli than auditory stimuli. The success of our emotional induction is evidenced by the PAM emotion labeling distributions: while the majority of participants identified the exact intended emotions, those who selected different labels consistently chose emotions adjacent to the target emotion in the circumplex plane (e.g., 'excited' for happiness scenario), demonstrating that the induced emotional states were accurately positioned in the intended valence-arousal space.

3.2. Statistical analysis

All statistical comparisons were performed using Friedman's test, followed by post-hoc pairwise comparisons with Bonferroni correction ($p < 0.05$). Normality was assessed using the Shapiro–Wilk test, and due to deviations from normality, a non-parametric approach was chosen.

Table 1 presents the mean and standard deviation of all features that were found to be significant in at least one comparison. Statistical results can be highlighted as follows:

ECG: RRTOTPWR was significantly higher in F compared to S, with a general trend of higher values associated with high-arousal emotions. This pattern was also observed for the LyapEx.

BVP: The ECG-BVP derived PAT showed particular relevance for F and S. These emotions exhibited mean values approximately 1 ms higher and lower, respectively, compared to the preceding neutral states, and differed significantly from each other. A trend of decreasing PAT with increasing arousal was observed.

GSR: Several GSR features distinguished F from all other emotions. Generally, GSR features showed lower values for low-arousal emotions (i.e., S and R) compared to high-arousal emotions (i.e., H and F). F consistently exhibited significantly higher values across all GSR features compared to the other three emotions. F was also consistently associated with higher GSR feature values compared to H in presence of high arousal. However, this relationship was not uniformly observed in the low-arousal category, where the relative magnitudes of R and S varied depending on the specific GSR feature under consideration.

RESP: Respiration-derived indices did not show statistically significant differences for any of the considered emotions.

3.3. Machine learning

Tables 2 and 3 present a comprehensive analysis of our classification results. Table 2 showcases the average F1-score, precision, recall, accuracy and the best test accuracy on the hidden test set computed for each model across the six feature selection methods. Conversely, Table 3 displays the average F1-score, precision, recall, and accuracy on the hidden test set computed and the best test accuracy for each feature selection method across all machine learning models employed. This dual perspective offers a nuanced understanding of classification performance, illuminating the relative strengths of both machine learning models and feature selection techniques.

Results further demonstrate that Logistic Regression (LR) consistently outperforms other models across all three classification tasks, achieving the highest mean test accuracies of 0.61±0.07, 0.72±0.05, and 0.80±0.05 for 4-emotion, arousal, and valence classification, respectively (see Table 2). LR also exhibits strong performance in other metrics, with notably high F1-scores (0.59±0.10, 0.70±0.06, 0.82±0.04) and balanced precision–recall pairs across tasks. For valence classification, LR achieves an exceptional recall of 0.90±0.05, indicating its effectiveness in identifying true positives. Linear Discriminant Analysis (LDA) demonstrates comparable performance, particularly in 4-emotion and valence classification. Among feature selection methods, the Square Method (SM) yields superior results with the highest mean test accuracies (0.53±0.10, 0.73±0.06, 0.74±0.09) across all tasks, while also maintaining strong F1-scores (0.51±0.12, 0.73±0.03, 0.74±0.08). SM also reveals balanced precision–recall metrics, particularly powerful in the arousal task (precision: 0.73±0.09, recall: 0.74±0.10). The K-Best selection method achieves competitive maximum accuracies (0.70, 0.80, 0.80), however with greater variability in its metrics and with the Friedman's test consistently displays showing the weakest performance of all the considered feature selection techniques (see Table 3). Beyond classification performance, we quantified variable importance by ranking each feature according to its selection frequency across the six employed feature selection methods. Fig. 7 presents a feature importance bubble diagram, offering a visual representation of this hierarchical analysis. Features were selected if identified as relevant by at least three out of the six feature selection methods employed, using this majority criterion as our selection

Table 1
Mean values and standard deviations of selected features across the 4 emotions. For clarity, only features showing one or more statistically significant differences, as detailed in the right-most column, are displayed. The last column shows the p-values for all comparisons, highlighting in bold the p-values related to significant differences (significance level 0.05) after applying the Bonferroni correction.

Emotions	Sadness (S)	Relaxation (R)	Happiness (H)	Fear (F)	*	p: S-R, S-H, S-F, R-H, R-F, H-F
RR TOTPWR [ms^2]	−115 (539)	89 (393)	172 (542)	265 (848)	F>S	0.46, 0.46, **0.02**, 1, 0.46, 0.46
LyapExp [s^{-1}]	0.027 (0.16)	−0.025 (0.16)	0.04 (0.13)	0.052 (0.16)	F>R	0.41, 0.99, 0.69, 0.35, **0.04**, 0.75
PAT [ms]	1.61 (3.79)	0.99 (2.69)	0.30 (3.2)	−1.85 (5.74)	S>F	0.97, 0.58, **0.01**, 0.84, 0.05, 0.30
GSR Amp peaks [nS]	0.59 (20.02)	4.20 (13.80)	5.21 (33.68)	22.43 (35.74)	F>S F>H	0.74, 1, **e−3**, 0.74, 0.07, **e−3**
GSR SD Amp peaks [nS]	−1.91 (17.65)	7.32 (18.43)	−2.89 (28.05)	15.69 (30.57)	F>S	0.19, 0.75, **e−3**, 0.75, 0.46, 0.07
AVGSR rise time [ms]	−67.20 (286.95)	−39.96 (226.19)	38.96 (220.20)	172.01 (274.29)	F>S	0.94, 0.74, **0.02**, 0.97, 0.09, 0.24
GSR n peaks	2.11 (7.31)	−0.92 (4.69)	3.36 (6.41)	7.97 (9.31)	F>S, F>R	0.89, 0.46, **e−3**, 0.13, **e−3**, 0.22
SDGSR der [nS]	0.25 (1.05)	0.25 (0.90)	0.93 (2.57)	1.41 (2.09)	F>S, F>R, F>H	0.88, 0.30, **e−5**, 0.75, **e−4**, **0.01**
Slope GSR [nS/s]	0.08 (0.27)	−0.04 (0.27)	0.05 (0.16)	0.18 (0.28)	F>R	0.05, 0.84, 0.40, 0.30, **e−3**, 0.08
AVGSR [nS]	93.51 (334.90)	67.08 (287.51)	256.02 (558.18)	458.84 (412.72)	F>S, F>R	0.52, 0.97, **0.03**, 0.26, **e−3**, 0.10
SDGSR [nS]	−23.07 (120.02)	12.98 (81.63)	2.55 (152.28)	138.36 (176.81)	F>S, F>R, F>H	0.88, 0.30, **e−5**, 0.74, **e−3**, **0.01**
Avg abs1 [nS]	0.33 (1.32)	0.22 (1.01)	0.83 (2.20)	2.00 (2.56)	F>S, F>R	0.99, 0.97, **0.01**, 0.92, **e−3**, 0.05
Avg abs2 [nS]	0.07 (0.31)	0.01 (0.19)	0.20 (0.45)	0.38 (0.59)	F>S, F>R, F>H	0.75, 0.99, **0.01**, 0.63, **e−4**, **0.02**
AVenv [nS]	3.54 (18.37)	1.83 (13.52)	11.40 (28.42)	22.87 (37.85)	F>S, F>R	0.99, 0.92, **e−3**, 0.84, **e−3**, 0.05

Table 2
Model Performance Across Feature Selection Methods. In the table are reported the mean ± std values of F1-score, Precision, Recall and Accuracy together with the maximum Accuracy on the hidden test set for each machine learning model, evaluated across six feature selection techniques.

4-emotion	LR	GNB	DT	LDA	SVM	RF	XGBoost
Mean Test F1-score	0.59 ± 0.10	0.46 ± 0.07	0.34 ± 0.03	0.57 ± 0.09	0.42 ± 0.08	0.42 ± 0.10	0.43 ± 0.04
Mean Test Precision	0.61 ± 0.13	0.55 ± 0.12	0.32 ± 0.04	0.61 ± 0.11	0.42 ± 0.11	0.41 ± 0.11	0.41 ± 0.04
Mean Test Recall	0.61 ± 0.07	0.51 ± 0.07	0.39 ± 0.03	0.58 ± 0.08	0.47 ± 0.05	0.45 ± 0.09	0.45 ± 0.04
Mean Test Accuracy	**0.61 ± 0.07**	0.50 ± 0.07	0.38 ± 0.03	0.58 ± 0.08	0.48 ± 0.05	0.44 ± 0.10	0.43 ± 0.04
Best Test Accuracy	**0.70**	0.60	0.45	**0.70**	0.55	0.50	0.50
Arousal							
Mean Test F1-score	0.70 ± 0.06	0.68 ± 0.03	0.63 ± 0.08	0.66 ± 0.04	0.69 ± 0.06	0.68 ± 0.06	0.59 ± 0.10
Mean Test Precision	0.75 ± 0.09	0.73 ± 0.04	0.60 ± 0.03	0.69 ± 0.04	0.77 ± 0.09	0.64 ± 0.06	0.60 ± 0.07
Mean Test Recall	0.66 ± 0.05	0.64 ± 0.05	0.67 ± 0.13	0.63 ± 0.04	0.63 ± 0.07	0.74 ± 0.10	0.59 ± 0.04
Mean Test Accuracy	**0.72 ± 0.05**	0.70 ± 0.02	0.62 ± 0.04	0.67 ± 0.03	**0.72 ± 0.05**	0.64 ± 0.05	0.60 ± 0.09
Best Test Accuracy	**0.80**	0.75	0.65	0.75	0.75	0.70	0.70
Valence							
Mean Test F1-score	0.82 ± 0.04	0.70 ± 0.02	0.61 ± 0.04	0.82 ± 0.04	0.74 ± 0.03	0.64 ± 0.08	0.64 ± 0.08
Mean Test Precision	0.75 ± 0.04	0.60 ± 0.03	0.62 ± 0.04	0.72 ± 0.06	0.69 ± 0.04	0.69 ± 0.10	0.65 ± 0.09
Mean Test Recall	0.90 ± 0.05	0.86 ± 0.05	0.65 ± 0.06	0.89 ± 0.05	0.80 ± 0.05	0.68 ± 0.10	0.70 ± 0.09
Mean Test Accuracy	**0.80 ± 0.05**	0.65 ± 0.03	0.64 ± 0.04	0.77 ± 0.06	0.73 ± 0.03	0.67 ± 0.06	0.65 ± 0.09
Best Test Accuracy	**0.85**	0.70	0.75	**0.85**	0.75	0.70	0.70

Table 3
Feature Selection Method Efficacy Across Models. In the table are reported the mean ± std values of F1-score, Precision, Recall and Accuracy together with the maximum Accuracy on the hidden test set for each feature selection method, evaluated across all implemented machine learning models.

4-emotion	Friedman's test	RFE	Model Based Ensemble	K-Best	MRMR	SM
Mean Test F1-score	0.39 ± 0.04	0.46 ± 0.07	0.44 ± 0.13	0.42 ± 0.15	0.50 ± 0.08	0.51 ± 0.12
Mean Test Precision	0.40 ± 0.10	0.48 ± 0.12	0.46 ± 0.13	0.43 ± 0.18	0.54 ± 0.10	0.55 ± 0.10
Mean Test Recall	0.44 ± 0.04	0.49 ± 0.07	0.48 ± 0.10	0.46 ± 0.14	0.53 ± 0.09	0.53 ± 0.10
Mean Test Accuracy	0.45 ± 0.04	0.44 ± 0.07	0.50 ± 0.15	0.48 ± 0.14	0.51 ± 0.09	**0.53 ± 0.10**
Best Test Accuracy	0.50	0.50	0.65	**0.70**	0.65	**0.70**
Arousal						
Mean Test F1-score	0.67 ± 0.04	0.67 ± 0.05	0.62 ± 0.06	0.66 ± 0.10	0.62 ± 0.08	0.73 ± 0.03
Mean Test Precision	0.67 ± 0.08	0.69 ± 0.10	0.66 ± 0.04	0.66 ± 0.10	0.68 ± 0.09	0.73 ± 0.09
Mean Test Recall	0.68 ± 0.07	0.66 ± 0.05	0.59 ± 0.09	0.66 ± 0.11	0.59 ± 0.12	0.74 ± 0.10
Mean Test Accuracy	0.67 ± 0.05	0.67 ± 0.08	0.65 ± 0.05	0.65 ± 0.09	0.66 ± 0.07	**0.73 ± 0.06**
Best Test Accuracy	0.75	0.75	0.70	**0.80**	0.75	**0.80**
Valence						
Mean Test F1-score	0.70 ± 0.08	0.69 ± 0.12	0.71 ± 0.10	0.71 ± 0.11	0.70 ± 0.09	0.74 ± 0.08
Mean Test Precision	0.67 ± 0.06	0.66 ± 0.11	0.65 ± 0.09	0.66 ± 0.06	0.64 ± 0.06	0.71 ± 0.07
Mean Test Recall	0.74 ± 0.17	0.74 ± 0.16	0.78 ± 0.13	0.78 ± 0.14	0.77 ± 0.15	0.80 ± 0.15
Mean Test Accuracy	0.67 ± 0.06	0.69 ± 0.12	0.70 ± 0.12	0.69 ± 0.08	0.68 ± 0.06	**0.74 ± 0.09**
Best Test Accuracy	0.75	0.80	**0.85**	0.80	0.80	**0.85**

threshold. The feature relevance diagram reveals seven significant variables (highlighted in green in Fig. 7) for valence and arousal classifications. Notably, n_peaks, was consistently selected by all six feature selection methods for both valence and arousal classifications, underscoring a robust predictive power. Among the top features for the arousal dimension, we find PAT and RR_TOTPWR in addition to n_peaks, while for valence, var is present alongside n_peaks. When comparing the same method for identifying the most relevant features, more features were selected as relevant for arousal (i.e., 23) compared to valence (i.e., 18). Considering that 7 features are shared between the two dimensions, we can rely on 34 of the most relevant features out of the 61 initially considered. To elucidate the decision-making processes underlying our models and interpret feature contributions, we applied the SHAP technique to the best-performing models coupled

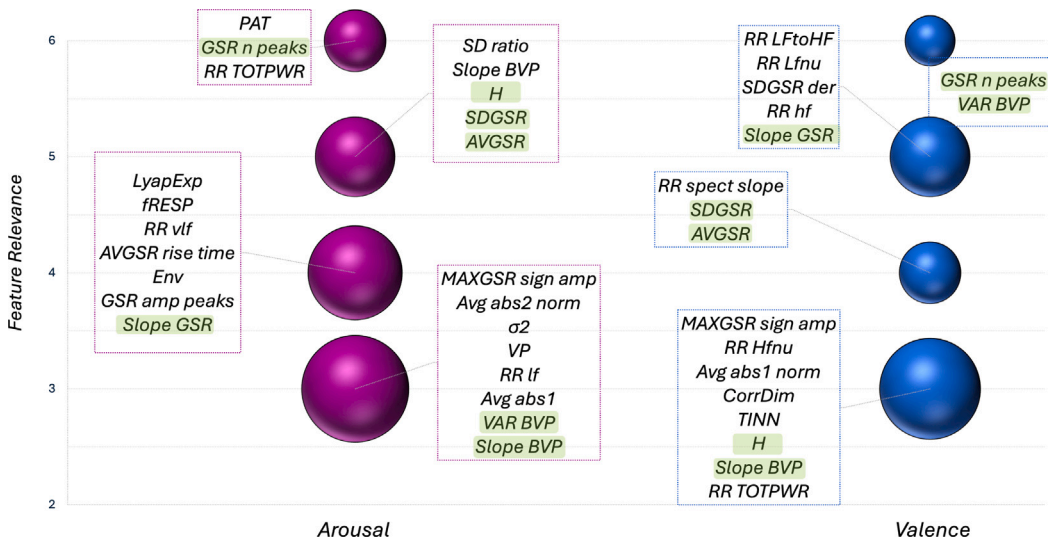

Fig. 7. Feature Relevance Bubble Diagram for Arousal (in violet) and Valence (in blue) Classification. Bubble size represents the number of variables with equivalent relevance scores. Green highlighted variables indicate relevance to both valence and arousal dimensions.

with the most effective feature selection methods, focusing on the LR models trained on datasets processed using the SM, which consistently demonstrated superior performance across both valence and arousal classification. As evident in Fig. 8, the SHAP method agrees with the analysis of feature relevance in Fig. 7. In particular, many of the features found in the bubbles corresponding to 6 and 5 feature relevance scores are also among the top features identified by the SHAP method, with n_peaks dominating both valence and arousal dimensions. The features highlighted in green in Fig. 7 are also present in both valence and arousal dimensions according to SHAP, and several features that are relevant for both valence and arousal dimensions in Fig. 8 (e.g. n_peaks, avg, sd, and slope_gsr, which exclusively pertain to the GSR signal) exhibit opposite trends in terms of their relative value ranges. For arousal, these features transition from blue to red moving from left to right, while the opposite occurs for valence. This suggests that, for arousal, high feature values are associated with high arousal levels whereas, for valence, high feature values are associated with low valence levels. Notably, SHAP analysis revealed VP as a highly relevant feature for arousal classification, in contrast to its lower importance shown in the bubble diagram. This feature is ranked 5th in importance for arousal by SHAP but is positioned lower in the bubbles chart.

Fig. 9 displays the separation between the four emotional states in a three-dimensional feature space defined by the top features identified by SHAP analysis: GSR n_peaks, RRLFtoHF ratio, and PAT. This visualization demonstrates the distinct physiological patterns associated with each emotion, with F (blue) characterized by lower PAT values and higher GSR peak counts, H (orange) showing moderate values across features, R (yellow) displaying elevated PAT and reduced GSR activity, and S (purple) exhibiting increased PAT and higher RRLFtoHF values.

4. Discussion

Our results provide a significant advancement in the field of emotion elicitation and recognition by use of VR technology. A key strength of the proposed research lies in the development of a novel VR protocol specifically designed to elicit emotions. To grant psychological validity, the experimental protocol was created in collaboration with a psychologist and accounting for the most influential scientific literature on the relationship between visual elements and emotional states. The protocol design incorporated specific visual cues such as angular shapes for high arousal and negative valence, rounded shapes for high valence and low arousal, as well as varying temporal stimuli speeds to modulate arousal levels. This multidisciplinary approach, combined with the use of custom-built VR scenarios, distinguishes our study from previous research that was exclusively relying on either pre-existing videos [26,27] or standard immersive virtual environments [69]. The meticulous design of our VR environments accounts for a range of visual and auditory cues, significantly enhances the ecological validity of the emotion elicitation process, potentially evoking more genuine and nuanced emotional responses compared to traditional methods [70].

Statistical analysis revealed significant differences separating emotional states for all signals except respiration, although only two key features were calculated from this variable. More in detail, ECG-derived RRTOTPWR and LyapEXP demonstrated statistically significant higher values in F compared to S and R, respectively, with an overall increasing trend with heightened arousal. Despite the absence of significant differences in RR LF and RR HF, RRTOTPWR dynamic changes revealed an initial surge in sympathetic activity, followed by rapid parasympathetic regulation to restore homeostasis [71] in presence of high-arousal emotional states such as F. The elevated LyapExp values observed in the F condition further suggested a more complex and less predictable HR pattern, also consistent with increased sympathetic activation and cardiovascular arousal associated with F states. Looking at ECG-BVP coupling, PAT was significantly lower in F compared to S, with a general decreasing trend from low to high arousal. F exhibited particularly low PAT values, with a mean among subjects lower than the preceding neutral state. Lower PAT values indicate heightened sympathetic control, manifesting as an acceleration of the pressure wave due to reduced transit time for blood to reach the periphery from the heart.

GSR emerged as the most effective signal for emotion discrimination, yielding the highest number of significant features. GSR features generally displayed a clear increasing trend for high-arousal emotions compared to low-arousal emotions. F consistently exhibited significantly higher values across all GSR features compared to the other three emotions. The discriminative power of GSR for arousal is well-documented in numerous studies [63] and is here confirmed with even stronger discriminating power. [64]. Interestingly, GSR features also demonstrated relevance for the valence dimension. Excluding features showing significant differences between F and R, likely due to substantial arousal differences, GSR Amp peaks, SDGSR der, SDGSR, and Avg abs2 exhibited significant differences between F and H - two high-arousal emotions with opposite valence. Notably, GSR Amp peaks showed differences only between F and H, not between F and R. While R had a lower mean than H, it also displayed a considerably reduced standard deviation compared to other emotions, suggesting a more

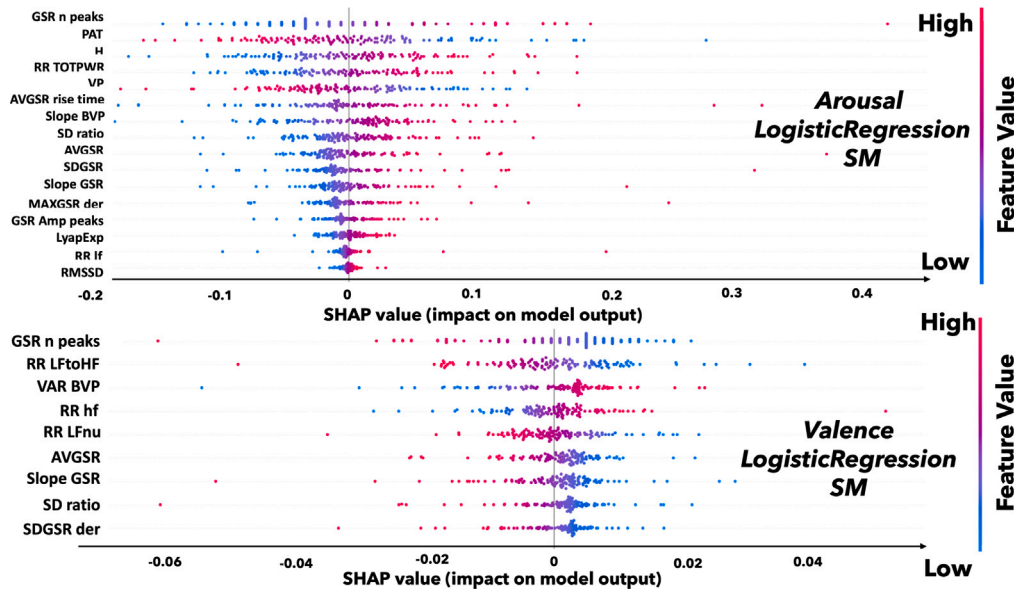

Fig. 8. SHAP summary plots for Logistic Regression models using SM feature selection in Arousal (up) and Valence (down) classifications. Features are ranked by their absolute SHAP values, indicating their overall importance in model predictions. Color represents the feature value (red: high, blue: low), while the horizontal axis shows the SHAP value impact on model output. This visualization reveals both the magnitude and direction of each feature's influence on classification decisions.

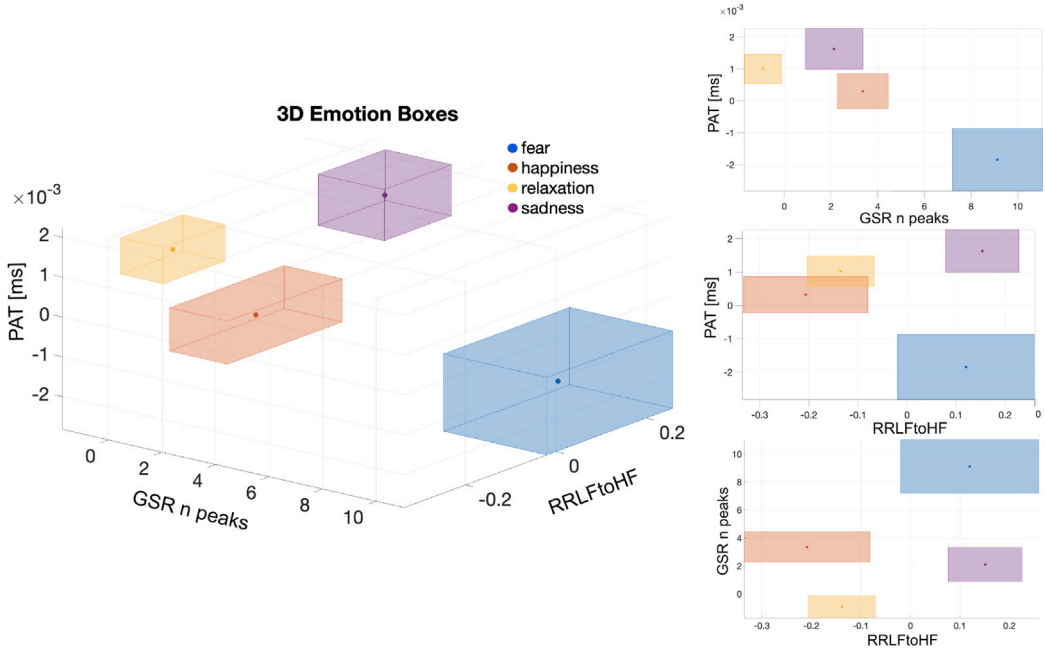

Fig. 9. 3D visualization of emotion clustering based on top physiological features. The center of each box represents the mean value of each feature (GSR n_peaks, RRLFtoHF, and PAT) for each emotion, with box dimensions extending to mean ± 2 standard errors. 2D projections on the right show feature pair relationships for clearer visualization of emotional state separation.

consistent response pattern in low-arousal states. While our analysis revealed significant differences between emotional states for ECG, BVP, and GSR signals, respiratory-derived indices did not show statistically significant differences between emotions. This finding yields several methodological considerations. First, we extracted only two respiratory features (respiratory frequency and mean breath amplitude), which may not capture the full complexity of breathing patterns during emotional responses. Second, respiratory dynamics typically operate on slower timescales compared to those observed through GSR and cardiovascular measures, potentially reducing their sensitivity to rapid emotional transitions in our protocol. Despite these limitations, our results align with some previous findings, suggesting that respiratory parameters may provide less discriminative information than other physiological signals in specific emotional contexts [72].

A closer examination of the classification results reveals that this study is able to achieve high accuracy in emotion recognition with a subject-independent approach. This is particularly noteworthy, as many previous studies have relied on subject-dependent approaches, which use data from a single subject at a time [73]. The ability to generalize across subjects demonstrates the robustness of the proposed methodology and its potential for real-world applications. The results presented in Tables 2 and 3 illustrate the effectiveness of the proposed methodology across various machine learning models and feature selection techniques. The LR model consistently outperforms other models,

with SM emerging as the most effective feature selection approach. It is interesting to note that SM, despite its simple implementation, proved to be superior among all commonly used feature selection methods. Our study demonstrates robust performance in emotion recognition, achieving an average 80% accuracy for arousal, 85% for valence, and 70% for four distinct emotions. These results are particularly noteworthy when contextualized within recent literature. For instance, [74] reported 77.5% accuracy for three emotional states (i.e., neutral, fear, sadness) in a subject-independent paradigm with 34 subjects, using cardiovascular, electrodermal, respiratory, and motor systems responses, with film clips as stimuli. [75] achieved 67.8% accuracy for valence and 77.0% for arousal using EEG (2 channels), eye data, and facial data with 49 subjects in a subject-independent approach, employing short videos and digital games as stimuli. An important previous study, [76], utilized 62-channel EEG data from 15 participants, with film clips as stimuli and three emotion classes. Here, authors evaluated both subject-dependent and subject-independent classifications. In the subject-dependent scenario, up to 91.07% accuracy was achieved using Graph regularized Extreme Learning Machine with differential entropy features. For subject-independent classification, the study reported an average accuracy of 60.93% using SVM with the same features. Extending our literature comparison to a broader context further substantiates the efficacy of our methodology. Other EEG-based methodologies have demonstrated variable efficacy: [77] reported 65% accuracy for quadrant-based affective classification employing wearable EEG sensors, which falls below our classification metrics. In dedicated valence discrimination studies, [78] attained approximately 65% accuracy using EEG in subject-independent paradigms, improving to 75% in subject-dependent analyses. This comparative analysis establishes that our approach achieves state-of-the-art performance while offering substantial practical advantages: non-invasive physiological monitoring compared to EEG-based methodologies, robust subject-independent generalizability addressing a persistent limitation in extant research, and compatibility with contemporary wearable technology platforms. These comparisons with the literature demonstrate our study's significant contribution to emotion recognition, particularly in achieving comparable performance with a less invasive setup. This advantage makes our method more suitable for real-world applications, where subject generalizability and non-invasive data collection are essential. Furthermore, with the growing prevalence of smartwatches and wearables capable of acquiring HRV and GSR signals [14], our approach shows promising potential for practical implementation. Moreover, this study demonstrates remarkable performance in valence classification, a dimension traditionally considered challenging to discriminate [18]. A significant contribution of this research is the comprehensive analysis of features through multiple feature selection methods and the application of XAI techniques. The utilization of SHAP provides valuable insights into the decision-making processes of our machine learning models, facilitating a deeper understanding of the physiological correlates of emotions. This interpretability is crucial for guiding future research and elucidating the often opaque nature of machine learning models in emotion recognition. As illustrated in Figs. 7 and 8, several features emerge as highly relevant for both valence and arousal dimensions, with some consistently selected across all feature selection methods. Notably, n_peaks demonstrates the highest efficacy, achieving an importance score of 6 for both valence and arousal, and ranking first in SHAP analysis for both dimensions. While the relationship between increased number of GSR peaks and higher arousal is well-established in literature [79], the connection between GSR peaks and valence has been less explored, with no clear evidence of a specific pattern. Our analysis suggests a potential role of GSR peaks in valence discrimination, contributing to the understanding of how GSR features might differently modulate arousal and valence dimensions of emotion. Consistent with our statistical analysis, top features for arousal include PAT and RRTOTPWR. Interestingly, VAR BVP, while not statistically significant, emerged as highly relevant for valence classification, with higher values associated with high arousal. In general, as introduced in Section 3, features related to the GSR signal that are deemed important for both classifications exhibit opposite trends in their values. Fig. 8 clearly shows that these features have high values for high arousal but also high values for low valence. This finding challenges the assumption of the orthogonality of the two dimensions, suggesting a need to reevaluate the independence of arousal and valence in the Continuous Model of Affect. It is noteworthy that simple features such as n_peaks, avg, sd, and slope_gsr rank among the most effective, despite their computational simplicity. A feature that has been highlighted as fairly important for arousal by all six feature selection methods, obtaining an importance score of 3, and ranked among the top five in importance by SHAP, is VP. This feature has been identified in other studies as significantly related to both arousal [34] and high task demand [80], warranting further investigation. The amplitude modulation of the BVP signal represents the volume of blood in the periphery; lower amplitudes are linked to higher peripheral blood pressure, which is associated with vasoconstriction. As shown in Fig. 8, along with PAT, VP is one of the few features where lower values are associated with high arousal (i.e., the color transitions from red to blue). Overall, our selected features demonstrate strong discriminative power across the four emotional states. As illustrated in Fig. 9, the 3D visualization of emotion clusters based on the three most important features selected by SHAP analysis across both valence and arousal dimensions (GSR n_peaks, RRLFtoHF, and PAT) reveals clear separation between F, H, R, and S. The distinct positioning of each emotion within this feature space provides visual confirmation of our quantitative findings, with minimal overlap between the emotion boxes even when accounting for variability (± 2 standard errors). This visualization reinforces our statistical analysis and helps explain the high classification accuracy achieved by our models, particularly for valence discrimination where sadness and fear are well-separated from happiness and relaxation along the RRLFtoHF and PAT dimensions.

The methodology and findings of this study demonstrate significant potential for practical applications. The high classification accuracy achieved with non-invasive physiological measures such as the one here used facilitates integration into contemporary wearable technology incorporating photoplethysmography and electrodermal activity sensors [81]. Implementation of our feature selection insights could enable continuous affective state monitoring in naturalistic settings. In clinical contexts, this system could augment mental health interventions by providing objective physiological correlates of emotional states to complement subjective assessments in anxiety and mood disorders. For individuals with psychiatric disorders, the system could function as a biofeedback mechanism for improving emotional awareness. Additionally, the high valence classification accuracy makes this approach valuable for neuromarketing research and user experience evaluation in digital environments where affective response discrimination is essential.

While the current study provides valuable insights into VR-based emotion elicitation and recognition, it is important to acknowledge its limitations and identify areas for future research. The controlled laboratory paradigm employed in this study, while ensuring methodological rigor and standardization, inherently constrains ecological validity. However, our VR-based approach represents an advancement over traditional flat-screen laboratory protocols where participants remain acutely aware of the laboratory environment. VR creates a more immersive experience, enhancing ecological validity compared to conventional emotion induction methods, though differences from real-world emotional experiences remain. The enhanced immersion effect provided by VR is further evidenced by the participants' perception of time during the experiment, with a reported average perceived duration of 34.43 ± 9.63 min versus the actual 42 min duration, suggesting sufficient engagement without boredom effects. Nevertheless, future research should extend validation to authentic real-world settings using

ecological momentary assessment methodologies, combining physiological measurements with in-situ self-reports or domain-specific validation in educational, clinical, or occupational contexts.

Another limitation is the focus on autonomic signals. Although informative, autonomic dynamics represent only a subset of the complex physiological responses associated with emotional experiences. To gain a more comprehensive understanding of emotions, future studies should consider integrating additional modalities, such as EEG to capture brain activity, facial expressions, eye tracking, and behavioral measures. This multimodal approach could provide a more holistic view of emotional states and potentially enhance the accuracy and robustness of emotion recognition algorithms. Another limitation is the sample size and diversity. While the current sample of 36 participants is comparable to many studies in the field, expanding the participant pool in terms of size, age range, gender balance, and cultural background would further validate the generalizability of the findings. Moreover, the current study focused on a basic set of emotions (sadness, relaxation, happiness, and fear). Future work could explore a broader range of emotions, including more complex and multifaceted affective states, such as awe, gratitude, or nostalgia. This would provide a more comprehensive understanding of the potential of VR in eliciting and recognizing a wider spectrum of emotions. Additionally, longitudinal studies could investigate the stability of emotional responses to VR stimuli over time and the potential long-term effects of VR-based interventions on emotional well-being. This could have important implications for the development of VR-based therapeutic interventions for mental health disorders. By addressing these limitations and exploring new avenues for research, future studies can build upon the foundation laid by the present work and contribute to a deeper understanding of the complex nature of emotions and the potential applications of VR technology in affective computing, mental health, and beyond.

Future implementation of this technology entails careful consideration of critical ethical aspects. Deployment of physiological-based emotion recognition systems requires rigorous protocols for informed consent acquisition, data privacy protection, mitigation of algorithmic biases, and contextual interpretation frameworks, particularly in sensitive domains such as clinical applications or educational assessment. These systems should function as complementary objective measures rather than autonomous diagnostic tools, with established safeguards for emotional data handling.

5. Conclusion

This study showcases the potential of VR in emotion elicitation and recognition using a novel protocol developed through interdisciplinary collaboration. The custom-built VR scenarios effectively induced genuine emotional responses, as evidenced by the statistical analysis of autonomic signals. Machine learning models, particularly LR with SM feature selection, achieved high accuracies in a subject-independent approach (85% for valence, 80% for arousal, and 70% for four distinct emotions), highlighting the robustness of the methodology.

XAI techniques provided valuable insights into the decision-making processes and the relevance of specific physiological features. GSR peaks emerged as the most significant feature for both dimensions, while other key features demonstrated distinct patterns across emotional states, with clear separation between emotions in the feature space defined by the number of GSR peaks, the ratio of low-frequency to high-frequency components in heart rate variability (RRLFtoHF), and Pulse Arrival Time (PAT, the time between the R-peak in the ECG and the corresponding pulse onset). Notably, our findings suggest that simple features such as GSR peaks and PAT offer strong discriminative power despite their computational simplicity, which has important implications for real-world implementation.

Our methodology offers several practical advantages over existing approaches: non-invasive physiological measurements compared to EEG-based systems, robust subject-independent generalizability, and compatibility with contemporary wearable technology. Furthermore, our careful VR design demonstrated high participant engagement with minimal discomfort, as evidenced by participants' underestimation of the experiment duration. These advantages position our approach favorably for practical applications in mental health contexts, user experience evaluation, and continuous emotion monitoring in naturalistic settings. In conclusion, despite limitations such as sample size and consideration and selective consideration of autonomic physiology, this study lays the groundwork for further advancements in VR-based emotion elicitation and recognition. Future research should explore the integration of multiple modalities, expand sample size, and investigate the application of this methodology to a broader spectrum of emotions and contexts, including specific clinical settings. Additionally, the relationship between GSR features and valence discrimination merits further investigation, as does the potential non-orthogonality of arousal and valence dimensions suggested by our feature analysis.

The integration of real-time emotion recognition using VR and wearable devices could open up new possibilities for affective computing and personalized interventions, while careful consideration of ethical dimensions and participant comfort will remain essential for responsible development and implementation.

CRediT authorship contribution statement

Edoardo Maria Polo: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation, Conceptualization. **Francesco Iacomi:** Writing – original draft, Software, Methodology, Data curation. **Alberto Valdes Rey:** Writing – original draft, Software, Methodology, Investigation, Data curation. **Davide Ferraris:** Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Alessia Paglialonga:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Riccardo Barbieri:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Ethics statement

The experimental protocol was conducted at the SpinLab facility of Politecnico di Milano, following written informed consent from all participants, and was approved by the Politecnico di Milano Research Ethics Committee (Approval No. 29/2021, date: 14/07/2021). All procedures were performed in compliance with relevant laws and institutional guidelines. The privacy rights of all human subjects were strictly observed throughout the study, and all data were collected and stored according to institutional data protection protocols. Participants were fully informed about the purpose of the research, the procedures involved, and their right to withdraw from the study at any time without consequences.

Funding

This work was supported in part by the PNRR Project PNRR-MAD-2022-12376716.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Riccardo Barbieri reports financial support was provided by Italian Ministry of Health. 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.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.110310.

References

[1] A. Caria, Hypothalamus, neuropeptides and socioemotional behavior, Brain Sci. 13 (9) (2023) 1303.

[2] K.R. Scherer, What are emotions? And how can they be measured? Soc. Sci. Inf. 44 (4) (2005) 695–729.

[3] P. Ekman, An argument for basic emotions, Cogn. Emot. 6 (3–4) (1992) 169–200.

[4] J. Posner, J.A. Russell, B.S. Peterson, The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology, Dev. Psychopathol. 17 (3) (2005) 715–734.

[5] M. Behnke, S.D. Kreibig, L.D. Kaczmarek, M. Assink, J.J. Gross, Autonomic nervous system activity during positive emotions: A meta-analytic review, Emot. Rev. 14 (2) (2022) 132–160.

[6] M. Egger, M. Ley, S. Hanke, Emotion recognition from physiological signal analysis: A review, Electron. Notes Theor. Comput. Sci. 343 (2019) 35–55.

[7] S.M. Hofmann, F. Klotzsche, A. Mariola, V. Nikulin, A. Villringer, M. Gaebler, Decoding subjective emotional arousal from EEG during an immersive virtual reality experience, Elife 10 (2021) e64812.

[8] L.A. Cattaneo, A.C. Franquillo, A. Grecucci, L. Beccia, V. Caretti, H. Dadomo, Is low heart rate variability associated with emotional dysregulation, psychopathological dimensions, and prefrontal dysfunctions? An integrative view, J. Pers. Med. 11 (9) (2021) 872.

[9] M. Nardelli, G. Valenza, A. Greco, A. Lanata, E.P. Scilingo, Recognizing emotions induced by affective sounds through heart rate variability, IEEE Trans. Affect. Comput. 6 (4) (2015) 385–394.

[10] M. Sklivanioti Greenfield, Y. Wang, M. Msghina, Similarities and differences in the induction and regulation of the negative emotions fear and disgust: a functional near infrared spectroscopy study, Scand. J. Psychol. 63 (6) (2022) 581–593.

[11] Q. Zhang, X. Chen, Q. Zhan, T. Yang, S. Xia, Respiration-based emotion recognition with deep learning, Comput. Ind. 92 (2017) 84–90.

[12] J.S. Heinisch, J. Kirchhoff, P. Busch, J. Wendt, O. von Stryk, K. David, Physiological data for affective computing in HRI with anthropomorphic service robots: the AFFECT-HRI data set, Sci. Data 11 (1) (2024) 333.

[13] A. Bellato, G. Sesso, A. Milone, G. Masi, S. Cortese, Systematic review and meta-analysis: Altered autonomic functioning in youths with emotional dysregulation, J. Am. Acad. Child Adolesc. Psychiatry 63 (2) (2024) 216–230.

[14] E. Mattern, R.R. Jackson, R. Doshmanziari, M. Dewitte, D. Varagnolo, S. Knorn, Emotion recognition from physiological signals collected with a wrist device and emotional recall, Bioengineering 10 (11) (2023) 1308.

[15] R.W. Picard, E. Vyzas, J. Healey, Toward machine emotional intelligence: Analysis of affective physiological state, IEEE Trans. Pattern Anal. Mach. Intell. 23 (10) (2001) 1175–1191.

[16] P.J. Lang, International Affective Picture System (IAPS): Technical Manual and Affective Ratings, The Center for Research in Psychophysiology, University of Florida, 1995.

[17] R.A. Stevenson, T.W. James, Affective auditory stimuli: Characterization of the international affective digitized sounds (IADS) by discrete emotional categories, Behav. Res. Methods 40 (1) (2008) 315–321.

[18] S. Koelstra, C. Muhl, M. Soleymani, J.S. Lee, A. Yazdani, T. Ebrahimi, et al., Deap: A database for emotion analysis; using physiological signals, IEEE Trans. Affect. Comput. 3 (1) (2011) 18–31.

[19] 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 (1) (2019) 196.

[20] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE Trans. Affect. Comput. 3 (1) (2011) 42–55.

[21] R. Somarathna, T. Bednarz, G. Mohammadi, Virtual reality for emotion elicitation–a review, IEEE Trans. Affect. Comput. (2022).

[22] M. Granato, D. Gadia, D. Maggiorini, L.A. Ripamonti, An empirical study of players' emotions in VR racing games based on a dataset of physiological data, Multimedia Tools Appl. 79 (45) (2020) 33657–33686.

[23] A. Felnhofer, O.D. Kothgassner, M. Schmidt, A.K. Heinzle, L. Beutl, H. Hlavacs, I. Kryspin-Exner, Is virtual reality emotionally arousing? Investigating five emotion inducing virtual park scenarios, Int. J. Hum.–Comput. Stud. 82 (2015) 48–56.

[24] B.O. Rothbaum, L. Hodges, R. Alarcon, D. Ready, F. Shahar, K. Graap, D. Baltzell, Virtual reality exposure therapy for PTSD Vietnam veterans: A case study, J. Trauma. Stress.: Off. Publ. Int. Soc. Trauma. Stress. Stud. 12 (2) (1999) 263–271.

[25] E. Pedroli, P. Padula, A. Guala, M.T. Meardi, G. Riva, G. Albani, A psychometric tool for a virtual reality rehabilitation approach for dyslexia, Comput. Math. Methods Med. 2017 (1) (2017) 7048676.

[26] E. Marqués Valderrama, M.A. Sarmiento Vega, I. Durán Díaz, J.A. Becerra González, I. Fondón García, Experimental setup and protocol for creating an EEG-signal database for emotion analysis using virtual reality scenarios, in: 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP, SciTePress, 2023, pp. 75–86.

[27] M. Li, J. Pan, Y. Li, Y. Gao, H. Qin, Y. Shen, Multimodal physiological analysis of impact of emotion on cognitive control in VR, IEEE Trans. Vis. Comput. Graphics 30 (5) (2024) 2044–2054.

[28] 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 (1) (2024) 132.

[29] N. Babu, U. Satija, J. Mathew, A.P. Vinod, Emotion recognition in virtual and non-virtual environments using EEG signals: Dataset and evaluation, Biomed. Signal Process. Control. 106 (2025) 107674.

[30] Q. Gong, X. Liu, Y. Ma, Real-time facial expression recognition based on image processing in virtual reality, Int. J. Comput. Intell. Syst. 18 (1) (2025) 8.

[31] F. Shaffer, Z.M. Meehan, C.L. Zerr, A critical review of ultra-short-term heart rate variability norms research, Front. Neurosci. 14 (2020) 594880.

[32] E.M. Polo, A.V. Rey, M. Mollura, A. Paglialonga, R. Barbieri, Exploring emotional responses in virtual reality through skin conductance signal, in: 2023 IEEE International Conference on Metrology for EXtended Reality, Artificial Intelligence and Neural Engineering, MetroXRAINE, IEEE, 2023, pp. 189–194.

[33] G. Valenza, L. Citi, A. Lanatá, E.P. Scilingo, R. Barbieri, Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics, Sci. Rep. 4 (1) (2014) 4998.

[34] E.M. Polo, A. Farabbi, M. Mollura, A. Paglialonga, L. Mainardi, R. Barbieri, Comparative assessment of physiological responses to emotional elicitation by auditory and visual stimuli, IEEE J. Transl. Eng. Heal. Med. 12 (2023) 171–181.

[35] A. Pinilla, J. Garcia, W. Raffe, J.N. Voigt-Antons, R.P. Spang, S. Möller, Affective visualization in virtual reality: An integrative review, Front. Virtual Real. 2 (2021) 630731.

[36] M.M. Bradley, P.J. Lang, The international affective picture system (IAPS) in the study of emotion and attention, 2007.

[37] M. Sabshin, Depression: Clinical, experimental and theoretical aspects, Arch. Gen. Psychiatry 19 (6) (1968) 766–767.

[38] M.R. Leary, Emotional responses to interpersonal rejection, Dialogues Clin. Neurosci. 17 (4) (2015) 435–441.

[39] S. Kaplan, The restorative benefits of nature: Toward an integrative framework, J. Environ. Psychol. 15 (3) (1995) 169–182.

[40] N. Javadzade, S.V. Esmaeili, V. Omranifard, F. Zargar, Effect of mindfulness-based stress reduction (MBSR) program on depression, emotion regulation, and sleep problems: A randomized controlled trial study on depressed elderly, BMC Public Health 24 (1) (2024) 271.

[41] M.E. Seligman, M. Csikszentmihalyi, Positive Psychology: An Introduction, vol. 55, no. 1, American Psychological Association, 2000, p. 5.

[42] C.L.M. Keyes, Social well-being, Soc. Psychol. Q. (1998) 121–140.

[43] A. Öhman, S. Mineka, Fears, phobias, and preparedness: toward an evolved module of fear and fear learning, Psychol Rev 108 (3) (2001) 483.

[44] R.N. Carleton, Into the unknown: A review and synthesis of contemporary models involving uncertainty, J. Anxiety Disord. 39 (2016) 30–43.

[45] S.F.M. Pizzoli, K. Mazzocco, S. Triberti, D. Monzani, M.L. Alcañiz Raya, G. Pravettoni, User-centered virtual reality for promoting relaxation: an innovative approach, Front. Psychol. 10 (2019) 479.

[46] A. Toet, M.van. Welie, J. Houtkamp, Is a dark virtual environment scary? CyberPsychology Behav. 12 (4) (2009) 363–371.

[47] P.M. Desmet, M.H. Vastenburg, N. Romero, Mood measurement with Pick-A-Mood: review of current methods and design of a pictorial self-report scale, J. Des. Res. 14 (3) (2016) 241–279.

[48] H. Sedghamiz, Matlab implementation of Pan Tompkins ECG QRS detector, 2014, Code Available at the File Exchange Site of MathWorks.

[49] D.J. Daley, D. Vere-Jones, An Introduction To the Theory of Point Processes: Volume I: Elementary Theory and Methods, Springer New York, 2003, pp. 211–275.

[50] R. Barbieri, E.C. Matten, A.A. Alabi, E.N. Brown, A point-process model of human heartbeat intervals: new definitions of heart rate and heart rate variability, Am. J. Physiol.-Hear. Circ. Physiol. 288 (1) (2005) H424–H435.

[51] S.D. Kreibig, Autonomic nervous system activity in emotion: A review, Biol. Psychol. 84 (3) (2010) 394–421.

[52] J.F. Thayer, F. Åhs, M. Fredrikson, J.J. Sollers III, T.D. Wager, (black). A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. Neuroscience & Biobehavioral Reviews, 2012, 36(2) 747–756.

[53] A. Haag, S. Goronzy, P. Schaich, J. Williams, Emotion recognition using bio-sensors: First steps towards an automatic system, in: Tutorial and Research Workshop on Affective Dialogue Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 36–48.

[54] E.H. Jang, B.J. Park, M.S. Park, S.H. Kim, J.H. Sohn, Analysis of physiological signals for recognition of boredom, pain, and surprise emotions, J. Physiol. Anthr. 34 (2015) 1–12.

[55] J. Bakker, M. Pechenizkiy, N. Sidorova, What's your current stress level? Detection of stress patterns from GSR sensor data, in: 2011 IEEE 11th International Conference on Data Mining Workshops, IEEE, 2011, pp. 573–580.
[56] C.L. Lisetti, F. Nasoz, Using noninvasive wearable computers to recognize human emotions from physiological signals, EURASIP J. Adv. Signal Process. 2004 (2004) 1–16.
[57] J. Fleureau, P. Guillotel, Q. Huynh-Thu, Physiological-based affect event detector for entertainment video applications, IEEE Trans. Affect. Comput. 3 (3) (2012) 379–385.
[58] J. Kim, E. André, Emotion recognition based on physiological changes in music listening, IEEE Trans. Pattern Anal. Mach. Intell. 30 (12) (2008) 2067–2083.
[59] C.A. Frantzidis, C. Bratsas, M.A. Klados, E. Konstantinidis, C.D. Lithari, A.B. Vivas, et al., On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications, IEEE Trans. Inf. Technol. Biomed. 14 (2) (2010) 309–318.
[60] L.R. Rabiner, B. Gold, Theory and Application of Digital Signal Processing:(by) Lawrence R. Rabiner (and) Bernard Gold, Prentice-Hall, 1975.
[61] I. Homma, Y. Masaoka, Breathing rhythms and emotions, Exp. Physiol. 93 (9) (2008) 1011–1021.
[62] P. Rainville, A. Bechara, N. Naqvi, A.R. Damasio, Basic emotions are associated with distinct patterns of cardiorespiratory activity, Int. J. Psychophysiol. 61 (1) (2006) 5–18.
[63] L. Juuse, D. Tamm, K. Lõo, J. Allik, K. Kreegipuu, Skin conductance response and habituation to emotional facial expressions and words, Acta Psychol. 251 (2024) 104573.
[64] M. Kołodziej, P. Tarnowski, A. Majkowski, R.J. Rak, Electrodermal activity measurements for detection of emotional arousal, Bull. Pol. Acad. Sci. Tech. Sci. 67 (4) (2019) 813–826.
[65] E.M. Polo, A. Farabbi, M. Mollura, L. Mainardi, R. Barbieri, Understanding the role of emotion in decision making process: using machine learning to analyze physiological responses to visual, auditory, and combined stimulation, Front. Hum. Neurosci. 17 (2024) 1286621.
[66] J. Gohumpu, M. Xue, Y. Bao, Emotion recognition with multi-modal peripheral physiological signals, Front. Comput. Sci. 5 (2023) 1264713.
[67] S. Akter, R.A. Prodhan, M.B. Mujib, M.A. Adnan, T.S. Pias, Evaluating the effectiveness of classification algorithms for EEG sentiment analysis, in: Sentiment Analysis and Deep Learning: Proceedings of ICSADL 2022, Springer Nature Singapore, Singapore, 2023, pp. 195–212.
[68] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (8) (2019) 832.
[69] J. Marín-Morales, J.L. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, E.P. Scilingo, et al., Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors, Sci. Rep. 8 (1) (2018) 13657.
[70] J. Kisker, T. Gruber, B. Schöne, Behavioral realism and lifelike psychophysiological responses in virtual reality by the example of a height exposure, Psychol. Res. 85 (2021) 68–81.
[71] J.F. Thayer, R.D. Lane, A model of neurovisceral integration in emotion regulation and dysregulation, J. Affect. Disord. 61 (3) (2000) 201–216.
[72] V. Kolodyazhniy, S.D. Kreibig, J.J. Gross, W.T. Roth, F.H. Wilhelm, An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions, Psychophysiology 48 (7) (2011) 908–922.
[73] Z. Lan, O. Sourina, L. Wang, Y. Liu, Real-time EEG-based emotion monitoring using stable features, Vis. Comput. 32 (3) (2016) 347–358.
[74] V. Kolodyazhniy, S.D. Kreibig, J.J. Gross, W.T. Roth, F.H. Wilhelm, An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions, Psychophysiology 48 (7) (2011) 908–922.
[75] W.K. Ngai, H. Xie, D. Zou, K.L. Chou, Emotion recognition based on convolutional neural networks and heterogeneous bio-signal data sources, Inf. Fusion 77 (2022) 107–117.
[76] W.L. Zheng, J.Y. Zhu, B.L. Lu, Identifying stable patterns over time for emotion recognition from EEG, IEEE Trans. Affect. Comput. 10 (3) (2017) 417–429.
[77] B. Nakisa, M.N. Rastgoo, D. Tjondronegoro, V. Chandran, Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors, Expert Syst. Appl. 93 (2018) 143–155.
[78] N. Jatupaiboon, S. Pan-Ngum, P. Israsena, Real-time EEG-based happiness detection system, Sci. World J. 2013 (1) (2013) 618649.
[79] F. D'Hondt, M. Lassonde, O. Collignon, A.S. Dubarry, M. Robert, S. Rigoulot, H. Sequeira, Early brain-body impact of emotional arousal, Front. Hum. Neurosci. 4 (2010) 1134.
[80] E.M. Polo, M. Mollura, A. Paglialonga, R. Barbieri, Listening effort: Cardiovascular investigation through the point process, in: 2022 Computing in Cardiology (CinC), vol. 498, IEEE, 2022, pp. 1–4.
[81] J.M. Peake, G. Kerr, J.P. Sullivan, A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations, Front. Physiol. 9 (2018) 743.