ORIGINAL ARTICLE

Real-time emotion recognition pipeline in videogames using physiological signals from a wearable device and facial action labels

Francisco Luna-Perejón · Miguel Civit · Luis Muñoz-Saavedra · Javier Civit-Masot · Manuel Domínguez-Morales · Lourdes Miró-Amarante

Received: 22 September 2025 / Accepted: 10 April 2026

Abstract

Advancing real-time emotion recognition in dynamic, naturalistic environments is pivotal for applications in gaming, mental health, and personalized education. This study introduces and evaluates an intelligent emotion recognition pipeline using physiological data from a wearable device, designed for this purpose. This system integrates continuous physiological sensing via the Empatica EmbracePlus wristband. Data were collected from 25 university students engaged in naturalistic gameplay across three commercial video games, forming the basis of the PaGER-Sync ADICVIDEO dataset. Emotions were labeled at a frame-level using FaceReader, providing high-granularity. The proposed classification pipeline employs a Random Forest classifier trained directly on short, lagged windows of raw physiological signals (EDA and BVP), eliminating the need for handcrafted feature extraction and achieving sub-50 ms inference latency, confirming suitability for real-time deployment. Addressing class imbalance with SMOTE-based data augmentation, the system achieved a robust performance, yielding 86.01% accuracy and a macro-averaged AUC of 0.98 across six discrete emotions. Model interpretability, crucial for building trustworthy intelligent systems, is quantified through SHAP-based analyses, revealing the individual contributions of physiological features. These compelling results demonstrate the feasibility of non-invasive, privacy-preserving, real-time emotion recognition using minimal wearable sensors, thereby supporting the development of scalable, interpretable, and computationally efficient affective computing applications in diverse real-world contexts.

Keywords Intelligent wearable systems · Emotion recognition · Physiological signals · Affective computing · Real-time monitoring · Machine learning · FACS

1 Introduction

Mental-health disorders affect nearly one billion people worldwide, and the World Health Organization (WHO) has called for scalable technologies that facilitate early detection, continuous monitoring, and timely intervention in everyday life [1]. Affective computing seeks to meet this need by enabling machines to recognise, interpret, and respond to human emotions, with applications ranging from digital therapeutics and stress prevention to adaptive learning environments [2, 3].

Beyond discrete taxonomies (e.g., joy, anger, fear), contemporary affective science often represents emotions in a continuous two-dimensional space spanned by *valence* (pleasant–unpleasant) and *arousal* (activation–calmness) [4]. This dimensional framework aligns well with physiological responses: sympathetic activation modulates skin conductance and cardiovascular dynamics along the arousal axis, while valence tends to reflect cortical and contextual appraisals [5].

Physiological signals offer objective, continuous, and largely involuntary markers of emotional states. Electrodermal activity (EDA) reflects sympathetic arousal, whereas blood-volume pulse (BVP), from which heart-rate variability (HRV) is derived, captures parasympathetic tone [5, 6]. These signals can be sampled unobtrusively using wrist-worn devices, enabling privacy-preserving monitoring in real-world settings [7]. Unlike facial expressions or voice, peripheral physiology cannot be easily masked or faked, making it a robust candidate for emotion inference.

While computer vision and speech-based models yield high performance in controlled settings, they present significant limitations: facial and vocal data are personally identifiable, require consistent lighting or silence, and impose ethical and technical barriers to real-world deployment [8, 9]. EEG and ECG offer high temporal fidelity but are intrusive and less practical for daily use [10, 11]. In contrast, wearable physiological sensing supports comfort, mobility, and on-device inference, which are essential for long-term, privacy-aware emotion monitoring.

Despite these advantages, emotion classification using EDA and BVP remains challenging due to weak signal-to-noise ratios, confounds from movement and temperature, inter-subject variability, and severe class imbalance in affective datasets [12]. Many studies rely on passive stimuli and post-hoc labels, which limit ecological validity. Robust real-time inference during interactive tasks remains underexplored.

There is a pressing need to evaluate whether minimal, non-invasive physiological signals alone can support accurate emotion recognition in ecologically valid, emotionally rich contexts such as gameplay. Most existing literature overlooks these scenarios, leaving open the question of whether wrist-worn devices can support real-time, privacy-conscious emotion monitoring in dynamic settings.

This study investigates whether two physiological signals–EDA and BVP–are sufficient to classify emotional states during commercial video game sessions. We leverage the PaGER-Sync ADICVIDEO dataset, which includes physiological recordings and facial annotations (via FaceReader) from 25 participants who played three commercial games designed to elicit diverse emotions. A preprocessing pipeline including artifact reduction, standardization, and SMOTE was applied for class balancing. Classical and deep learning models were trained and evaluated and SHAP analysis was used to interpret feature contributions.

Research Objective: To develop a privacy-preserving, wearable-based system using only EDA and BVP achieve accurate and interpretable emotion recognition during interactive gameplay, without relying on audiovisual or self-report modalities.

Research Questions:

- **RQ1:** Are two peripheral signals (EDA and BVP) sufficient to develop a robust emotion recognition in dynamic, interactive settings?
- **RQ2:** Can the individual contributions of physiological features be quantified using SHAP-based model explanation techniques, supporting transparent intelligent systems?
- **RQ3:** Can the pipeline for emotion classification be computationally efficient and suitable for real-time deployment on wearable devices?

The remainder of the manuscript is organized as follows. Section 2 reviews related work on physiological emotion recognition; Sect. 3 details the dataset, exploratory analysis, preprocessing steps and modeling techniques; Sect. 4 presents empirical findings; Sect. 5 discusses their implications; and Sect. 6 advises on the possible limitations of the study.

2 Related work

Over the past decade, both classical machine learning (ML) and deep learning (DL) have propelled affect recognition from physiological data [7, 13–16]. Classical models such as Random Forests (RF), logistic regression, support-vector machines (SVM), and *k*-nearest neighbors (KNN) rely on hand-crafted temporal, frequency-domain and non-linear features selected by domain experts [13, 14]. For example, Domínguez-Jiménez et al. [14] extracted BVP and EDA features from 37 participants and reached *100%* accuracy on a three-emotion task (amusement, sadness, neutral) using RF-RFE and SVM, underscoring the discriminative power of EDA when the class distribution is carefully balanced. Muñoz-Saavedra et al. [17] went further by designing an inexpensive wrist-worn device that records only GSR (EDA) and PPG. Optimizing a three-layer MLP, they exceeded 92% precision across valence–arousal levels for 22 subjects, demonstrating that a bichannel wearable can rival more complex systems with rigorous feature selection and tuning.

DL models–multi-layer perceptrons (MLPs), convolutional neural networks (CNNs), and long short-term memory (LSTM) networks–learn hierarchical representations directly from raw signals [13, 15]. Pidgeon et al. [13] applied a 1-D CNN to peripheral channels from the DEAP corpus, reaching 62–64% accuracy for binary arousal/valence classification with isolated channels, highlighting that single-channel DL approaches face performance ceilings without fusion or pre-engineering. Qing et al. [16] proposed a stacked auto-encoder with ensemble voting, achieving 63% EEG-based emotion recognition on DEAP.

Combining heterogeneous biosignals can boost performance at the cost of complexity and portability. The DEAP study by Koelstra et al. [10] used EEG, peripheral signals, and audiovisual fusion to slightly improve arousal/valence classification. Schmidt et al. [7] released the WESAD corpus (chest and wrist sensors), achieving up to 93% stress detection with Random Forests. Indrasiri et al. [18] applied a multi-scale attention LSTM across multiple biosignal domains in VR, outperforming single-domain models but requiring three concurrent sensors. Still, recent evidence [17] suggests that well-designed bichannel wristband solutions can rival complex setups with over 92% accuracy.

Public datasets like DEAP [10], WESAD [7], DREAMER [11], AMIGOS, and EEVR support reproducibility and cover various stimulus modalities. However, most rely on passive stimuli and self-reports, limiting temporal resolution and ecological validity, these are gaps our gameplay-based dataset aims to address.

Real-world affective data are inherently imbalanced, with mild and neutral states overrepresented and intense emotions underrepresented [15, 19]. This skews decision boundaries and reduces recall for minority classes [20, 21]. SMOTE is a standard oversampling technique that interpolates new synthetic samples in feature space and improves classifier robustness [22]. Recent work explores GAN-based augmentation [19], but it requires large corpora and extensive tuning. In this study, SMOTE is selected for its balance between effectiveness and computational simplicity. Despite progress in physiological emotion recognition, most prior research relies on either multimodal setups that compromise ecological validity, or simplified laboratory protocols that fail to capture the dynamics of real-world interaction. Furthermore, few studies explicitly address the joint challenges of motion artifacts, imbalanced class distributions, and temporal synchronization of physiological signals with emotion labels in naturalistic environments.

In contrast, our work explores whether robust, frame-level emotion classification is possible using only two peripheral signals–EDA and BVP–recorded from a wrist-worn device during real-time video gameplay. By combining continuous physiological monitoring, facial-based emotion annotation, and gameplay as a dynamic affective stimulus, this study provides one of the first end-to-end evaluations of wearable-based affect recognition in an ecologically valid, interactive context.

3 Methods

3.1 System overview

The proposed emotion recognition system is designed as an end-to-end pipeline that integrates wearable physiological sensing, automated emotion labeling, feature engineering, and supervised classification. This modular architecture ensures reproducibility, scalability, and real-time deployment potential for affective computing applications in naturalistic settings.

3.2 Data collection protocol

This study builds upon data collected within the ADICVIDEO research project (PID2022-141172OA-I00), which aims to develop robust emotion recognition systems based on peripheral physiology. The resulting dataset, named **Pager-Sync ADICVIDEO** [23] (see Sect. 6) (*Player Affective Gaming Experience & responses - Synchronized Dataset*), comprises multimodal recordings from 25 university students during standardized gameplay sessions involving three commercial video games–*Tetris*, *Sonic Racing*, and *Fall Guys*. These games were selected for their accessibility and their capacity to elicit a broad spectrum of emotional responses while remaining intuitive and accessible to participants with minimal gaming experience.

Physiological signals were continuously recorded using the Empatica EmbracePlus wristband (Empatica Inc., Boston, MA, USA) [24], a research-grade wearable device offering high-resolution streams of electrodermal activity (EDA, in μS), blood volume pulse (BVP, in μV), skin temperature (°C), triaxial acceleration (g), and heart rate (bpm). Data for each sensor were collected at their respective device-native sampling rates in a natural gaming setup: EDA at 16 Hz, BVP at 64 Hz, and Skin Temperature at 1 Hz.

In parallel, video footage of participants was captured throughout the gameplay session. These recordings were analyzed with FaceReader 9 (Noldus Information Technology) [25], a validated computer vision platform that implements the Facial Action Coding System (FACS) to infer emotional states based on detected Action Units (AUs). For each video frame (30 fps), the software outputs the likelihood of six basic emotions: *neutral*, *angry*, *happy*, *sad*, *surprised*, and *disgusted*. The emotion with the highest probability at each frame was selected as the ground-truth label for supervised classification.

The comprehensive study protocol received explicit approval from the Research Ethics Committee of the University of Seville, under reference number **2024-1858**, with official approval granted in September 2024. Data were collected within the ADICVIDEO research project (PID2022-141172OA-I00), which focuses on emotion recognition based on physiological responses during gameplay.

3.3 Exploratory feature analysis

To guide feature selection and better understand the underlying variance structure of the data, we applied Principal Component Analysis (PCA). This exploratory step aimed to reveal dominant axes of physiological variation across participants, potentially uncovering patterns useful for both outlier detection and clustering. Critically, this step served as an exploratory data analysis tool to identify physiological outliers and understand the variance structure of the dataset prior to supervised modeling, ensuring that the training data represented a consistent physiological baseline.

By projecting the data onto uncorrelated principal components (PCs) ordered by explained variance [26], PCA guide model interpretability, training stability, and generalization [27]. Furthermore, PCA projections allow for visualization of participant clustering, class separability, and detection of potential outliers [28]. Analyzing component loadings also sheds light on which signal characteristics contribute the most to variance, supporting informed feature engineering. Although limited to linear relationships, PCA offers a robust and interpretable foundation for the analysis of physiological data in the early stages [2].

Following the PCA, K-Means clustering was applied to the first two principal components (PC1 and PC2) to identify natural physiological subgroups among participants. This unsupervised clustering aimed to group individuals based on their dominant patterns of physiological responses. A total of three distinct clusters (k=3) were identified, which helped in understanding the inter-individual variability and detecting outlier physiological profiles within the study cohort.

3.4 Feature selection and preprocessing

Based on exploratory analysis, feature selection focused on EDA and BVP signals due to their strong, well-documented links to autonomic nervous system (ANS) activity and emotional arousal [5, 6].

EDA directly measures sympathetic arousal via sweat gland activity, capturing both tonic (baseline) and phasic (stimulus-evoked) responses. BVP enables estimation of heart rate variability (HRV), which reflects the balance between sympathetic and parasympathetic modulation. Stress and high-arousal emotions are typically associated with reduced HRV and increased sympathetic tone [29].

Although skin temperature can reflect ANS modulation (e.g., vasoconstriction during stress), its response is slower and more influenced by environmental conditions [30, 31]. For this reason, temperature was excluded from the primary feature set to focus on signals offering more direct, rapid sensitivity to emotional state changes.

For modeling, each data instance was constructed by concatenating raw physiological values from time-lagged windows. Specifically, values spanning from $t - 4$ to $t + 4$ for both BVP and EDA were extracted relative to each timestamp, resulting in 18 raw data points (9 BVP and 9 EDA samples) per instance, representing a local physiological context. Data were then split into training (80%) and test (20%) subsets. To strictly avoid data leakage and overfitting, a subject-independent split was employed where distinct participants were held out for the test set. This ensures that the model is evaluated on its ability to generalize to new users, rather than simply recalling patterns from seen subjects. The resulting distribution is shown in Table 1. As can be observed, there are notable class imbalances. To address this inherent class imbalance, which is typical in naturalistic emotion data and can severely bias classifiers towards majority classes, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training data [32]. SMOTE systematically synthesizes new samples for minority classes by interpolating between existing minority class instances and their nearest neighbors in the feature space. This process effectively balances the class distribution, thereby improving the model's ability to learn and generalize from underrepresented emotional states and mitigating majority class bias without simply duplicating existing data points. Table 1 also includes the results of applying SMOTE over the training subset; each minority class was oversampled to match the majority class (Angry), resulting in exactly 250,801 samples per category and a total balanced training set of 1,504,806 samples. To robustly evaluate model performance and optimize hyperparameters, a k-fold cross-validation approach was applied to this augmented training data. This internal validation ensured the model's generalization capabilities before final evaluation on the unseen test set..

Table 1 Number of samples per emotion class in the original, training, SMOTE-augmented training, and test sets

Emotion class	Total samples	Training set (80%)	Training set after SMOTE	Test set (20%)
Neutral	304,867	243,893	250,801	60,974
Angry	313,500	250,801	250,801	62,701
Happy	58,854	47,083	250,801	11,771
Sad	17,972	14,377	250,801	3,595
Surprised	7,513	6,010	250,801	1,503
Disgusted	321	256	250,801	65
Total	**703,028**	**562,419**	**1,504,806**	**140,609**

 https://doi.org/10.1007/s00521-026-12307-5

3.5 Model portfolio and hyperparameter optimization

To benchmark supervised emotion recognition strategies, five machine learning algorithms were selected on the basis of their complementary representational capacities and their proven suitability for sequential physiological data analysis [33–35]. The primary objective was to determine the best-performing approach for integration into a wearable based system. Although the final system is intended to operate in real-time, inference may be executed either on-device or on a higher-performance edge/server node; consequently, both lightweight and moderately complex models were retained in the study.

All models were trained in the SMOTE balanced training subset introduced in Sect. 3.5. For each algorithm, grid search combined with 5-fold cross-validation was used to exhaustively explore the hyperparameter spaces listed in Table 2. Iterative learners (neural networks and gradient-boosting trees) were limited to 200 optimization iterations with an early-stopping criterion of patience = 15, halting training when the mean validation loss failed to improve for 15 consecutive epochs.

The following models were considered:

Multi-Layer Perceptrons (MLPs). Fully connected feed-forward networks with ReLU activations and a Softmax output layer for multiclass prediction. Architectures ranging from three to four hidden layers (topology up to 256 neurons) were examined. Adam optimization was used and regularization of dropout (0.15–0.30) mitigated overfitting [33].

1-D Convolutional Neural Networks (CNNs). Temporal convolutional stacks captured short-term dependencies in the BVP and EDA sequences, followed by a global average pooling stage and a dense classifier [36]. Variants differed in the number of convolutional blocks (one or two), filter counts (32–96), kernel sizes (3 or 5) and dropout levels (0.4–0.5).

Random Forest. An ensemble of decision trees that aggregates class votes to reduce variance. Grid search varied the number of trees (50–200), maximum depth (10–30 or unlimited) and node-splitting constraints, enabling a balance between accuracy and inference latency.

LightGBM. Gradient-boosted decision trees optimized for speed and memory efficiency. The grid encompassed 50–200 boosting rounds, depths up to 30, learning rates from 0.05 to 0.15, and 15 or 31 terminal leaves [34].

Table 2 Hyperparameter values considered for grid search

Model	Hyperparameters	Considered values
Multi Layer Perceptron	Layers and nodes	[128, 64, 32], [256, 64, 32], [128, 128, 64, 32], [256, 128, 64, 32]
	Batch size	32, 64, 96
	Learning rate	0.0001, 0.0005, 0.0015
	Dropout per layer	0.15, 0.2, 0.3
Convolutional Neural Network (+1 Hidden Dense Layer)	CNN layers and kernels	[32], [64], [32, 64], [64, 96]
	kernel size per layer	3, 5
	Hidden dense layer nodes	32, 64
	Batch size	32, 64, 96
	Learning rate	0.0001, 0.0005, 0.0015
	Dropout per layer	0.4, 0.5
Random Forest	Number of estimators (trees)	50, 100, 150, 200
	Maximum depth	10, 20, 30, No limit
	Minimum samples for split	2, 6, 10
	Minimum samples for leaf	1, 2, 4
LightGBM	Number of estimators (trees)	50, 100, 150, 200
	Maximum depth	10, 20, 30, No limit
	Learning rate	0.05, 0.1, 0.15
	Number of leaves	15, 31
Logistic Regression	Solver	lbfgs, saga
	Regularization coefficient (C)	0.1, 1, 10

Multinomial Logistic Regression. A linear baseline that applies the Softmax function to weighted feature sums. Solver type (LBFGS or SAGA) and regularization strength ($C = 0.1$–10) were tuned. Its interpretability and minimal computational footprint make it a valuable reference [35].

The best configuration for each model, defined as the parameter set that maximizes the mean validation accuracy across the five folds, was retained for the final evaluation on the hold-out test set. All grid-search and cross-validation runs were executed on a Google Colab virtual machine equipped with two Intel Xeon CPU cores at 2.20GHz and 12 GB of RAM, operating entirely on CPU resources; consequently, the training and latency figures reported later correspond to a reproducible commodity-hardware setting. The complete hyper-parameter grids are provided in Table 2 to facilitate replication of the study.

3.6 Model evaluation framework

To assess classifier performance under real-world conditions, we designed an evaluation protocol that emphasizes both overall accuracy and sensitivity to minority emotion classes. This is particularly important given the strong class imbalance in the dataset, with emotions like *Neutral* and *Angry* dominating the label distribution, and others like *Disgusted*, *Sad*, or *Surprised* being underrepresented.

All models were trained and evaluated using SMOTE-augmented training sets to improve class balance. Subsequently, all models underwent hyperparameter tuning using cross-validation as described in Section 3.5. Following this optimization, the final optimized models were trained and then evaluated on a dedicated, unseen test set. Performance was measured across multiple standard and custom metrics:

- **Accuracy** – the proportion of correctly classified instances across all classes.
- **Macro-Averaged AUC** – the average Area Under the Curve for each class in a one-vs-rest setting, providing a balanced view of discrimination performance across classes.
- **Precision, Recall, and F1-Score** – calculated per class to assess how well each emotional state was detected.
- **Confusion Matrices** – used to visualize misclassifications and reveal systematic confusion between emotionally similar states.
- **ROC Curves** – plotted for each model to examine multiclass separability and sensitivity.

In addition to these standard metrics, we defined a custom evaluation measure, the *Combined Score*, tailored to our use case. This metric was designed to reward balanced classification, particularly for minority emotions that are critical for system responsiveness. It is computed as:

$$\begin{aligned}\text{Combined Score} = \text{Overall Accuracy} \\ + \text{Avg\_Mid\_Recall} \\ + 0.5 \times \text{Recall\_Disgusted}\end{aligned} \tag{1}$$

where:

$$\text{Avg\_Mid\_Recall} = \frac{\text{Recall\_Sad} + \text{Recall\_Surprised}}{2}.$$

This formulation prioritizes:

- **General performance** (accuracy),
- **Sensitivity to moderately rare classes** (*Sad*, *Surprised*),
- and **Specific emphasis on the most underrepresented class** (*Disgusted*).

All classifiers were compared using this Combined Score alongside the standard metrics to provide a holistic view of performance, ensuring that models selected for deployment are not only accurate overall, but also capable of reliably detecting less frequent yet emotionally salient states.

3.6.1 Best model interpretability and real-time inference

To support interpretability of the best-performing model, we employed SHAP (SHapley Additive exPlanations) [37], a unified framework for explaining predictions of any machine learning model. SHAP computes the contribution of each input feature to the model's output by estimating Shapley values from cooperative game theory. This allows both global analysis–identifying the most important features across the dataset–and local analysis–explaining individual predictions. Given its solid theoretical grounding and suitability for tree-based classifiers, SHAP was chosen to increase trust, transparency, and insight into how specific EDA and BVP patterns influence emotional state predictions.

Inference latency was assessed in the same Google Colab environment used for training. The virtual machine provided two Intel Xeon CPU cores running at 2.20GHz and 12GB of RAM under a KVM hypervisor. No dedicated GPU acceleration was used. Mean wall-clock time was measured for each batch-size condition from 1 to 6 samples. For every condition, 12 000 forward passes were executed, the first 100 runs were discarded to remove initialisation bias, and the remaining times were averaged. Per-batch and per-sample latencies are reported together with 95% confidence intervals obtained by bootstrapping. This procedure quantifies whether the selected model satisfies the sub-100 ms response target for interactive systems and, by exploiting small batches, approaches the 33 ms frame interval required at 30 fps for real-time wearable or edge deployment.

4 Results

4.1 Principal Component Analysis (PCA) results

We conducted a Principal Component Analysis (PCA) on participant-level summary metrics derived from the Empatica EmbracePlus recordings. For each participant, statistical descriptors including mean, standard deviation, minimum, and range were extracted from electrodermal activity (EDA), blood volume pulse (BVP), and skin temperature signals. This dimensionality reduction step aimed to uncover dominant physiological axes of inter-individual variation in a naturalistic gaming context.

The first two principal components (PC1 and PC2) collectively explained approximately 63% of the total variance (PC1: 41.7%, PC2: 21.2%).

Table 3 summarizes the PCA loadings for PC1 and PC2, ordered by their overall vector magnitude, highlighting the relative contribution of BVP, EDA, and temperature features to each component.

As shown in Fig. 1 (Panel B), the variable loadings plot clearly reveals the underlying structure of the principal components. **PC1**, which accounts for 41.7% of the variance, is largely defined by opposing physiological responses:

- Strong positive loadings are observed for `BVP_mean` (0.451) and `BVP_std` (0.431), indicating that higher mean and variability in blood volume pulse contribute significantly to the positive end of PC1. Notably,

Table 3 PCA loadings (PC1 and PC2) ordered by overall magnitude

Variable	PC1	PC2	Magnitude
bvp_min	**−0.429**	−0.399	0.586
bvp_std	**0.431**	0.395	0.585
bvp_mean	**0.451**	0.327	0.557
eda_mean	**−0.338**	**0.421**	0.540
eda_std	**−0.289**	**0.386**	0.483
temp_range	−0.297	**0.369**	0.474
temp_min	0.324	−0.279	0.428
eda_min	−0.184	0.172	0.252
temp_mean	−0.042	0.089	0.098

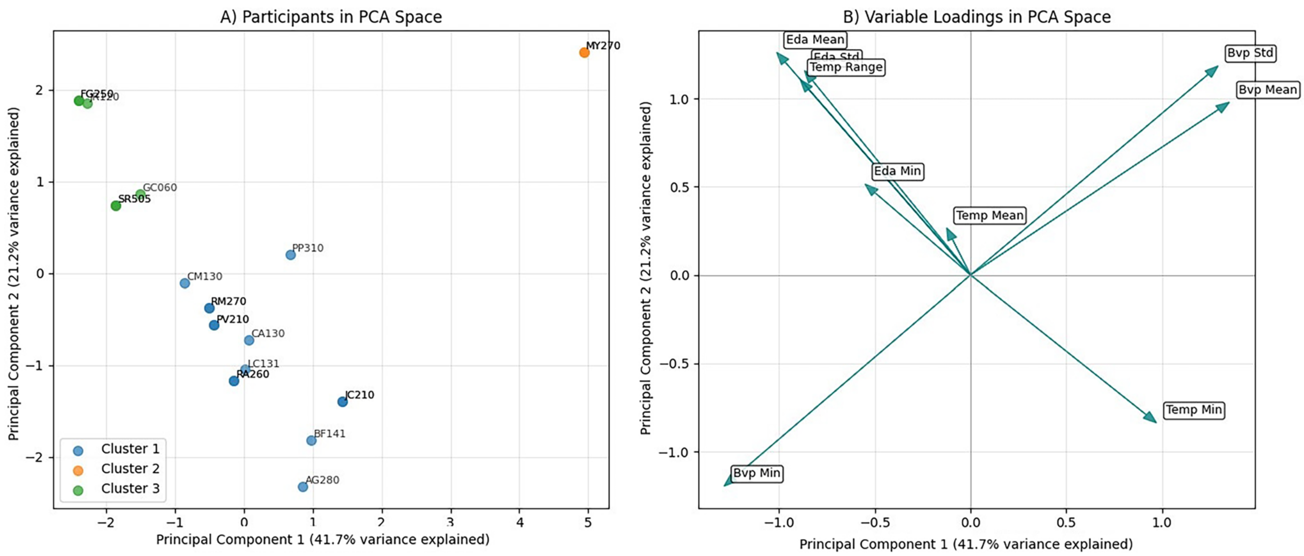


Fig. 1 PCA analysis of participant-level physiological metrics: (**a**) Participant projections onto the first two principal components–accounting for 41.7% and 21.2% of total variance, respectively–with markers colored by k-means cluster assignment. (**b**) Variable loadings plot illustrating the contribution of Electrodermal Activity (EDA), Blood Volume Pulse (BVP), and Skin Temperature features to the latent dimensions of inter-individual variability

`bvp_min` loads strongly and negatively (−0.429), suggesting that participants with higher BVP mean/std tend to have lower minimum BVP values within their recorded range.

- Conversely, `EDA_mean` (−0.338) and `EDA_std` (−0.289) load negatively on PC1, implying an inverse relationship where elevated electrodermal activity contributes to the negative end of PC1.

It is notable that while PC1 and PC2 together explain 63% of the variance, the remaining 37% likely reflects participant-specific noise or higher-order physiological interactions not captured in these axes. This highlights the complexity of autonomic responses in real-world gaming contexts, while still supporting the utility of EDA and BVP as dominant, interpretable sources of variation. Thus, these data give us confidence in using this specific set of physiological signals (EDA and BVP) to train the model for emotion recognition.

Figure 1 (Panel A) visually represents the projection of individual participants onto the PC1–PC2 space, color-coded by the k-means clustering results that identified three distinct groups based on their physiological profiles:

- **Cluster 1 (blue)**: This is the most numerous group, predominantly located in the lower-left to mid-right portion of the graph. These participants show a characteristic physiological profile with variable levels of BVP activity and relatively lower contributions on the PC2 axis (EDA/Temperature). They represent the most common response pattern in our cohort.
- **Cluster 2 (orange)**: This cluster consists of a single notable participant, 'MY270', prominently situated in the upper-right quadrant of the graph (high scores on PC1 and PC2). This positioning indicates a unique physiological profile characterized by **very high BVP values (mean and standard deviation), along with high EDA activity and variability, and a wide temperature range**. 'MY270' is clearly identified as an *outlier* with markedly elevated physiological reactivity compared to the rest of the sample.
- **Cluster 3 (green)**: A smaller group of participants ('FG250', 'GC060', 'SR505') is located in the upper-left quadrant (high scores on PC2 and negative on PC1). This profile is distinguished by **high EDA activity (mean and standard deviation) and an elevated temperature range**, with BVP contributions more towards the negative extreme of PC1, suggesting a lower average cardiovascular reactivity compared to Cluster 2.

The main separation of these clusters along PC1 reinforces that the distinction between the dominance of BVP activity and EDA activity is a key differentiator among participants, while PC2 adds a layer of variability related to thermoregulation and baseline EDA.

Importantly, these PCA-derived axes emerge from data collected in an unscripted, naturalistic gaming context, reinforcing the robustness of EDA and BVP as key physiological channels for emotion recognition even outside laboratory-controlled stimuli.

4.2 Classification performance overview

This section details the performance of the machine learning models evaluated for emotion recognition based on physiological signals. Performance metrics, including accuracy, macro-averaged AUC, per-class recall, and the custom Combined Score designed to reward balanced classification, are presented (see Sect. 3 for details). We also analyze class-specific errors and discrimination capacity using ROC curves and confusion matrices. The analysis is structured into four subcomponents: overall model comparison, ROC analysis, confusion patterns, and model interpretability.

4.2.1 Model effectiveness results

To assess the effectiveness of different classification strategies, all models were evaluated under consistent conditions, including data preprocessing, feature scaling, and SMOTE-based class balancing. Performance was assessed using multiple metrics, including overall accuracy, macro-averaged AUC, per-class recall, and Combined Score (see Sect. 3). Table 4 summarizes the results for the five classifiers evaluated.

The **Random Forest** model achieved the highest overall performance, with an accuracy of 0.8601 and a macro-averaged AUC of 0.9845. It also maintained strong recall on minority classes, Sad (0.905), Surprised (0.954), and Disgusted (0.875), showing that the classifier generalizes well despite the original label imbalance. The Combined Score of 2.227 confirms that the model excels not only in global accuracy but also in detecting infrequent, but psychologically salient emotional states. His performance was achieved with a forest of 150 decision trees (estimators), each limited to a depth of 30 levels maximum, and restricted to split only when at least 10 samples are present in a node and to keep a minimum of 2 samples in every leaf settings identified as optimal during the hyperparameter search.

LightGBM, another ensemble-based model, also benefited from SMOTE balancing and showed competitive recall for Surprised (0.925) and Sad (0.759), but its overall accuracy of 0.5110 (51.1%) and macro AUC (0.8803) were substantially lower than those of Random Forest. This suggests that, although boosting methods can perform well in tabular data, Random Forests may offer greater robustness in small-to-medium datasets with high class imbalance.

Deep learning models, including Multi-Layer Perceptrons (MLPs) and 1D Convolutional Neural Networks (CNNs), achieved moderate performance. The best MLP configuration (4 layers, 128 units) yielded 0.4573 (45.7%) accuracy and 0.8397 AUC, with decent recall for Sad (0.698) and Surprised (0.731). The CNN model

Table 4 Performance comparison of evaluated classification models with all emotion recalls

Model	Acc.	Rec. (Neu.)	Rec. (Ang.)	Rec. (Hap.)	Rec. (Sad.)	Rec. (Sur.)	Rec. (Dis.)	AUC	Score
Random Forest	0.8601	0.782	0.924	0.897	0.905	0.954	0.875	0.9845	2.227
LightGBM	0.5110	0.307	0.640	0.749	0.759	0.925	0.922	0.8803	1.810
MLP 4L (256)	0.4573	0.207	0.564	0.739	0.698	0.731	0.641	0.8397	1.490
CNN 2 Blocks	0.3931	0.153	0.549	0.682	0.568	0.921	0.906	0.8186	1.500
LogReg	0.2130	0.150	0.250	0.180	0.449	0.046	0.100	0.6551	0.550

The columns reflect the recall rates for the emotional categories of neutral, angry, happy, sad, surprised, and disgusted

(2 convolutional blocks) performed slightly worse in accuracy of 0.3931 (39.3%), but maintained high recall for Surprised (0.921), indicating some ability to capture local temporal patterns in the data. However, both architectures struggled to generalize reliably, likely due to limited data availability and model sensitivity to class imbalance.

Finally, the **multinomial logistic regression** baseline showed the weakest performance (accuracy: 0.2130 (21.3%), AUC: 0.6551), and extremely low recall for minority emotions such as Surprised (0.046). This confirms the limitations of linear classifiers in capturing the non-linear relationships present in physiological-emotional dynamics, particularly in an imbalanced multiclass setting.

Taken together, these results underscore the importance of model choice and class balancing strategies. Ensemble methods, particularly Random Forests, emerged as the most suitable candidates for emotion classification from peripheral physiological signals, offering both high accuracy and balanced recall across emotional states critical for real-time affective computing applications.

4.2.2 ROC curve analysis

Figure 2 presents the ROC curves for the evaluated models, illustrating their multiclass discrimination capability. As shown in Fig. 2(a), the Random Forest with SMOTE demonstrates superior and consistent AUC values across all emotion classes, confirming its robustness in handling class imbalance and diverse emotional states. LightGBM (Fig. 2(b)) also shows reasonable performance, though with lower overall separation. In contrast, MLP (Fig. 2(c)) and CNN models (Fig. 2(d)) display lower and more variable AUC curves, particularly struggling to maintain sensitivity for minority classes such as Disgusted or Sad. These results highlight the advantage of ensemble-based methods in this imbalanced, multiclass emotion recognition task.

4.2.3 Confusion matrix analysis

Figure 3 shows the normalized confusion matrices for the main models. The Random Forest model (Fig. 3(a)) achieves balanced classification, with high true positive rates across most emotion categories and comparatively fewer misclassifications. LightGBM (Fig. 3(b)) also performs well but exhibits more confusion between emotionally adjacent classes such as Neutral and Happy. In contrast, MLP (Fig. 3(c)) and CNN models (Fig. 3(d)) show pronounced confusion, particularly missclassifying minority emotions like Disgusted and Surprised into majority classes such as Neutral or Angry. These matrices illustrate the critical role of both data augmentation and model choice in achieving robust multiclass emotion classification from imbalanced physiological data.

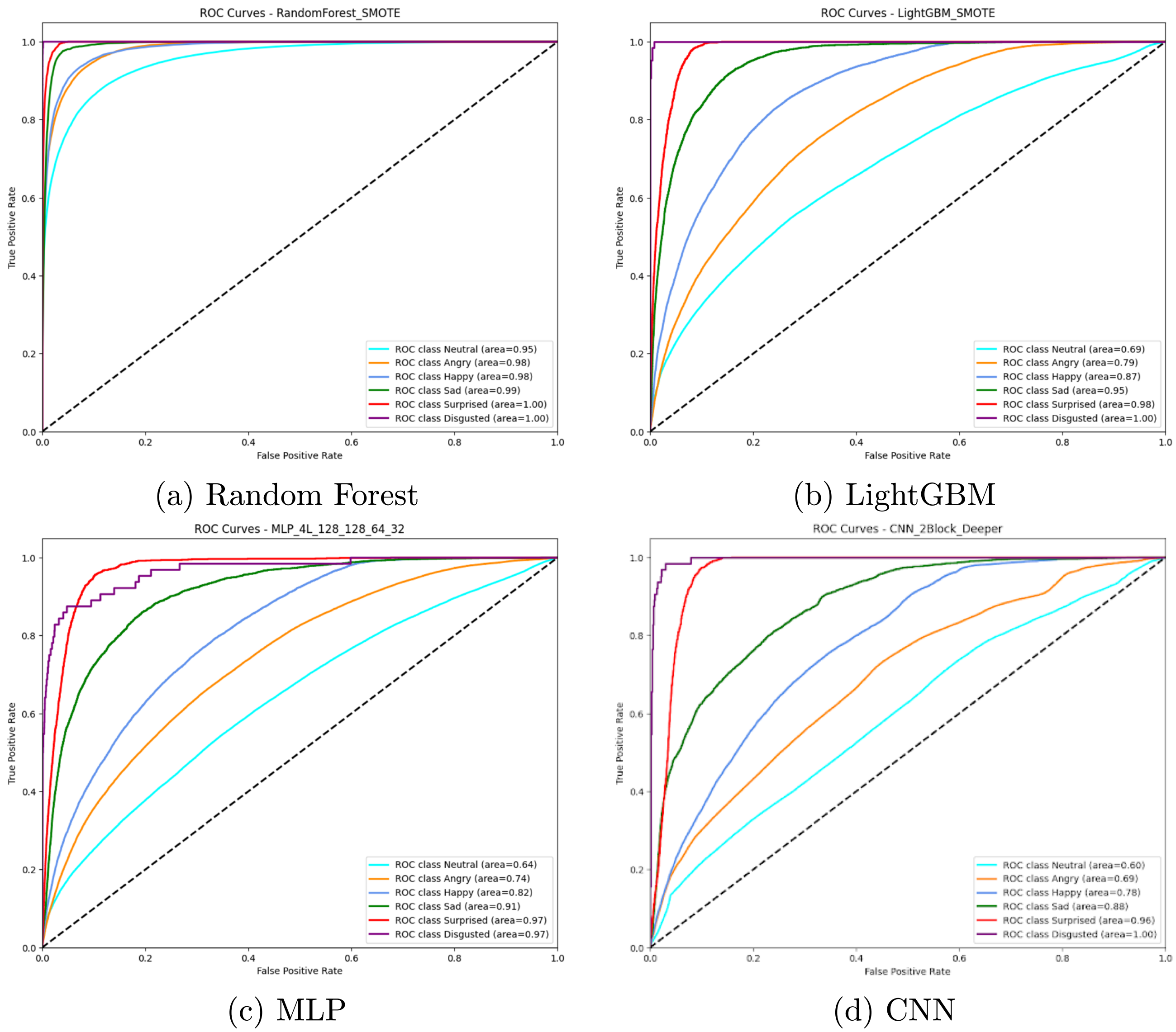
4.3 Interpretability and model inference latency

To better understand how physiological features influence emotion prediction, we applied SHapley Additive exPlanations (SHAP) to the best-performing Random Forest classifier. SHAP assigns an importance value to each feature based on its marginal contribution to model predictions, enabling both global and local interpretability [37].

Figure 4a presents a SHAP summary plot for all features in the model. Electrodermal activity (EDA) features, particularly those at future and current time offsets (e.g., `EDA_t+4`, `EDA_t+0`), show the highest average impact on the classifier's output across all emotion categories. In contrast, blood volume pulse (BVP) features contribute significantly less, reinforcing the greater discriminative power of skin conductance signals in our setting.

In addition, Fig. 4b visualizes a local force plot for a representative instance, illustrating how specific EDA values drive the model toward or away from particular predictions. Here, high EDA activity at `t+4` and `t-2` strongly push the output toward the target class.

These insights validate our feature selection strategy and highlight the value of SHAP for debugging and interpreting wearable-based affect recognition models in real-world scenarios.



(a) Random Forest (b) LightGBM (c) MLP (d) CNN

Fig. 2 ROC curves for the evaluated models, illustrating multiclass performance. (**a**) Random Forest shows superior, consistent AUC across classes. (**b**) LightGBM, (**c**) MLP, and (**d**) CNN show varying performance, with Random Forest highlighting its robustness in this imbalanced, multiclass emotion recognition task

Inference latency was evaluated on the same Google Colab host by varying the batch size from one to six samples (Table 5). With a single-sample batch is obtaining a mean latency of 44.06 ms (95% CI 43.89–44.27 ms). Increasing the batch size exploited the two CPU threads more efficiently. At a batch of two samples, the model produced one prediction every 22.49 ms on average while keeping the batch turnaround below 45 ms. This is well inside the 66-ms window required to process the physiological signals corresponding to two consecutive video frames at 30 fps, meaning that real-time operation becomes feasible even without additional optimization. Batches of three or four samples further reduced the per-sample latency to 14.95 ms and 11.53 ms, respectively, offering additional headroom for communication and preprocessing overheads. However, the absolute batch latency remained below 47 ms in all sizes tested, indicating that moderate buffering does not sacrifice interactive responsiveness.

Despite these gains, the Random Forest model–150 trees with maximum depth of 30–still occupies tens of megabytes in memory, a footprint beyond the capacity of typical microcontroller-class wearables. Practical

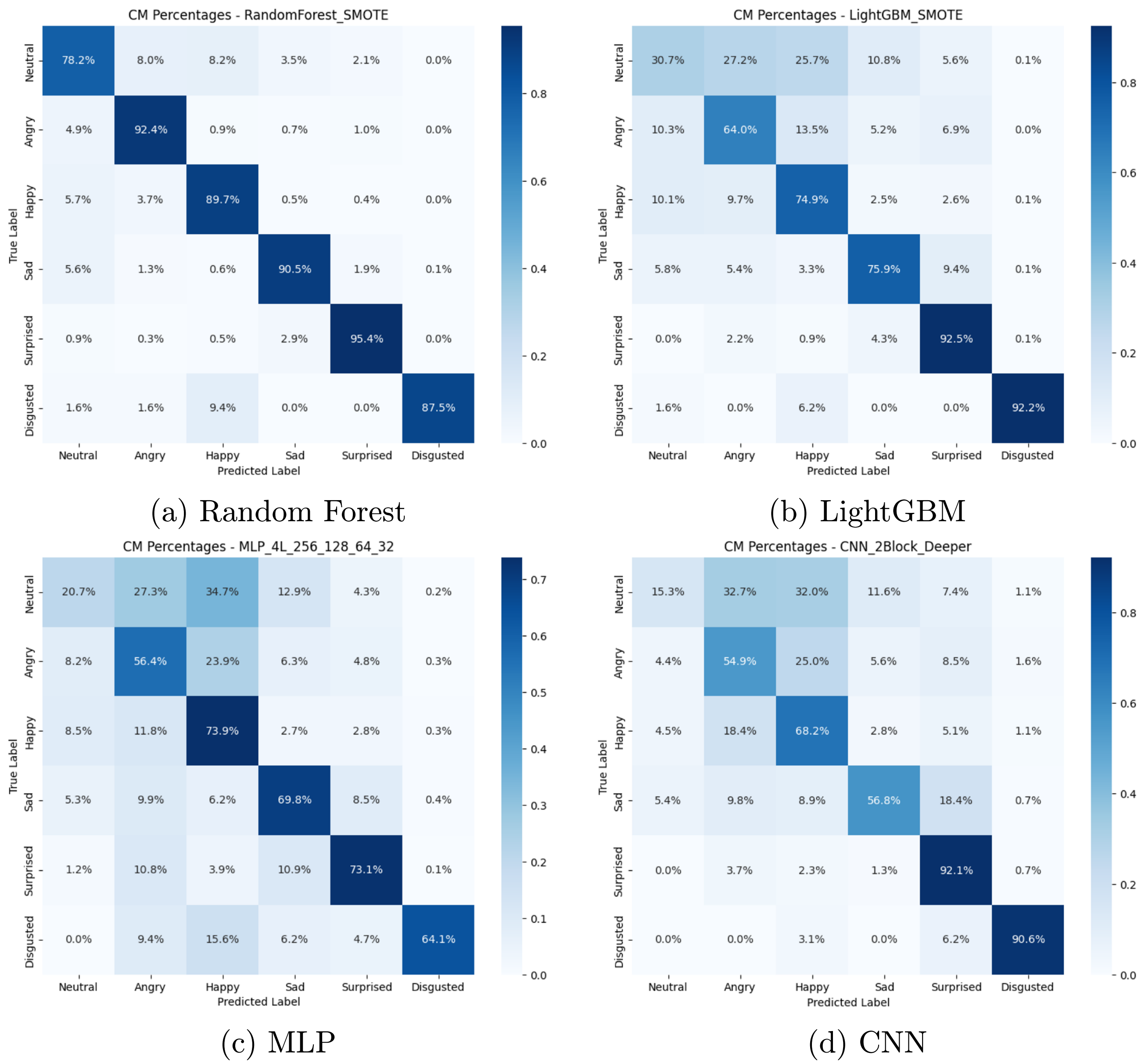
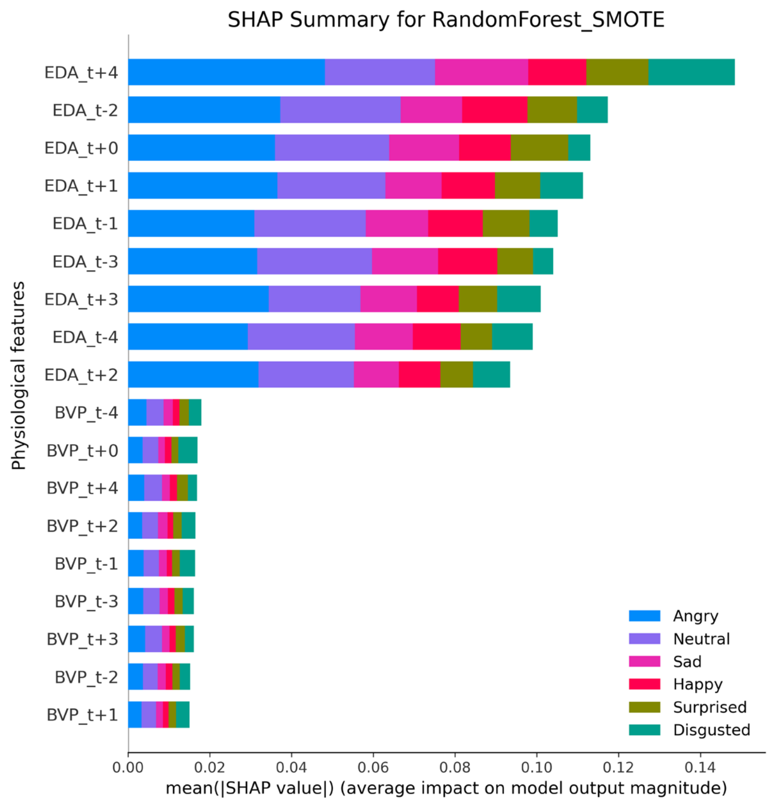


Fig. 3 Normalized confusion matrices of evaluated models. (**a**) Random Forest demonstrates balanced classification across emotion classes. (**b**) LightGBM, (**c**) MLP, and (**d**) CNN show higher confusion, especially among minority emotional states

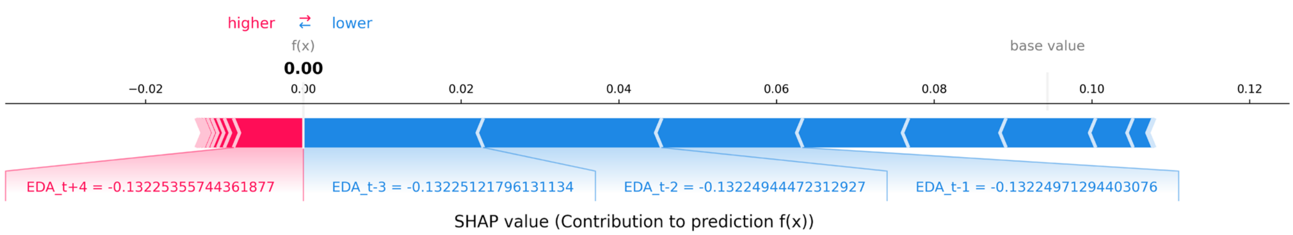
deployment therefore favors either offloading the sensor stream to a nearby edge device with resources comparable to the Colab configuration or employing model-compression and multithreaded execution on more capable embedded processors. Future work will benchmark such optimizations against the latency targets established here.

In any case, these findings, combined with the SHAP-based feature analysis presented earlier, support the model's physiological plausibility and improve interpretability, enhancing its suitability for affective computing applications.

 https://doi.org/10.1007/s00521-026-12307-5



(a) SHAP summary plot with X axis displaying the mean absolute impact of physiological features on model output magnitude across all emotional classes and Y axis corresponding to names of physiological signals in different times.



(b) Individual force plot showing the transition from the base value to the specific output $f(x) = 0.00$ for a single instance.

Fig. 4 Global (**a**) and local (**b**) interpretability analysis for the Random Forest classifier using SHAP values. (a) Features related to Electrodermal Activity (EDA) lags exhibit the highest global importance, significantly outweighing Blood Volume Pulse (BVP) metrics. (**b**) For a specific prediction, the model shifts from the expected base value of 0.094 toward zero, primarily driven by the negative contributions (blue segments) of historical EDA lags ($t-1$, $t-2$, and $t-3$), which act as the dominant predictors for this classification

Table 5 Random-Forest inference latency on CPU (Google Colab) as a function of batch size

Batch size	Mean per batch (ms)	Mean per sample (ms)	CI low (ms)	CI high (ms)	Speed-up ×
1	44.06	44.06	43.89	44.27	1.00
2	44.98	22.49	22.41	22.57	1.96
3	44.86	14.95	14.90	15.01	2.95
4	46.11	11.53	11.42	11.69	3.82
5	45.74	9.15	9.11	9.19	4.82
6	46.29	7.72	7.68	7.75	5.71

CI: 95% bootstrap confidence interval

Table 6 Comparative studies on emotion recognition using physiological signals. Studies are ordered chronologically to illustrate the evolution of methods

Authors (Year)	Dataset (n; context)	Physiological signals	Classification problem	Feature extraction	ML/DL models	Accuracy
Schmidt et al. (2018) [7]	WESAD (15); baseline, stress (TSST), amusement, meditation	ECG, EDA, EMG, Resp., Temp.	Stress detection (binary); 3-class: baseline/stress/amusement	Yes	Random Forest, SVM	93% (binary); 80% (3-class)
Domínguez-Jiménez et al. (2020) [14]	Own dataset (37); video-clip elicitation	PPG, GSR	Three discrete emotions: amusement, sadness, neutral	Yes	RF-RFE + SVM	100%
Pidgeon et al. (2022) [13]	DEAP (32); music-video viewing	BVP, EDA	Two levels: valence and arousal	Yes	1-D CNN	62–64% (valence)
Muñoz-Saavedra et al. (2023) [17]	Own dataset (22); video-clip elicitation	GSR (EDA), PPG	Three levels of valence and arousal (low/mid/high)	Yes	Optimised MLP	92–98% (valence)
Indrasiri et al. (2024) [18]	Own dataset (23); immersive 360º VR videos	EEG, ECG, EDA, BVP, Temp.	Seven levels of valence and arousal	No	LSTM + Multi-Scale Attention	83.5% (valence)
This Study (2025)	PaGER-Sync ADICVIDEO (24); naturalistic gameplay; FaceReader labels	BVP, EDA	Six basic emotions (frame-level annotation)	No	Random Forest + SMOTE	86.01%

5 Discussion and conclusions

Our findings demonstrate that peripheral physiological signals collected from the Empatica EmbracePlus wristband–specifically BVP and EDA–are sufficient to classify discrete emotional states with high accuracy in a naturalistic setting. Among the evaluated models, Random Forest with SMOTE-based data augmentation achieved the best performance (86.01% accuracy; macro-averaged AUC of 0.98). This underscores the importance of ensemble methods and effective class balancing when handling real-world emotional data with inherent imbalance.

5.1 Discussion

Table 6 resumes six representative investigations spanning 2018–2025 and highlights how emotion-recognition pipelines have evolved from multi-sensor laboratory setups to more minimalist, wearable solutions. The earliest wearable-centred work, WESAD [7], employed a chest-plus-wrist platform with five physiological channels and achieved 93 % accuracy for binary stress detection, although performance dropped to 80 % when a third class (amusement) was introduced. Subsequent studies narrowed signal requirements while boosting performance in constrained scenarios: Domínguez-Jiménez *et al.* [14] used only PPG and GSR yet reported a remarkably high (100 %) three-class accuracy, albeit on a highly controlled, video-elicitation dataset.

Research has also explored end-to-end deep learning on single channels. Pidgeon et al. [13] trained a 1D CNN on raw BVP and EDA from DEAP, reaching 62–64 % for valence prediction. This pointed towards evidence that peripheral channels alone face ceiling effects without careful feature engineering or the use of more complex architectures. Addressing this, Muñoz-Saavedra et al. [17] optimised an MLP on a custom two-signal wrist device and surpassed 92 % accuracy across three arousal/valence levels, indicating that careful windowing and hyper-parameter tuning can compensate for limited sensor breadth.

Systems with more complex architectures continue to raise absolute scores: Indrasiri et al. [18] combined EEG, ECG and five peripheral channels with a multi-scale attention LSTM, attaining 83.5 % across seven valence/arousal bins in an immersive VR setting–though at the cost of three concurrent devices, high bandwidth and continuous audiovisual capture.

In comparison with these previous works, our system offers a set of advantages. First, our system only relies on two signals (BVP and EDA) captured through a small wrist-worn device, yet it tackles a markedly harder six-emotion classification problem that includes infrequent categories such as *surprise* and *disgust*. Secondly, the proposed pipeline operates without the need for a handcrafted feature extraction; the model ingests short, logged windows of the raw signals, allowing the Random-Forest ensemble to learn discriminative patterns directly. This approach eliminates the need of feature engineering processes while maintaining high performance. Third, emotion labels are generated automatically and at frame-level resolution with FaceReader. This helps remove inter-subject bias differences while providing a pipeline where automatic emotion labeling can easily be updated according to a reference system.

Moreover, even under these constraints, the proposed system still achieves 86.01 % accuracy and 0.98 macro-AUC in a naturalistic gameplay scenario–showing that a scalable, privacy-preserving affect monitoring is feasible without multimodal fusion, bulky sensors, or manual feature engineering.

Another critical distinction lies in ground truth labeling. Many benchmark datasets rely on post-task self-reports (e.g., DEAP, WESAD), introducing memory bias and limiting temporal resolution. In contrast, our study employs frame-level facial expression annotation using FaceReader during gameplay sessions, providing consistent, automated labeling with high temporal granularity. Moreover, the integrating a step of automated labeling in the pipeline means that, for a set of data (such as captured video and physiological signals from gameplay sessions), these labels could be easily updated if the FaceReader software was significantly improved or a different one was preferred.

Furthermore, our source for emotion elicitation differs from traditional lab-based protocols that use static images or scripted video clips. The gaming context offers dynamic, interactive experiences that elicit authentic, varied emotional responses, enhancing ecological validity and supporting real-world applications such as adaptive gaming, mental health monitoring, or personalized user experiences.

Despite using only two physiological signals, our system achieves comparable or superior accuracy relative to both more complex, multimodal systems and to similar systems with bigger sensor arrays. Its simplicity and portability make it suitable for scalable applications in digital health, education, and affective human-computer interaction.

Future work will explore integrating additional signals available from the Empatica Embrace PLus (e.g., skin temperature, accelerometry) and applying sequential deep learning models such as LSTMs or temporal CNNs to capture emotional transitions over time. As well as the possibility of further improving its accuracy with minimal yet automatic feature engineering. We also aim to examine the complementarity between FaceReader-derived annotations and subjective self-reports, further validating the robustness of emotion labeling in naturalistic settings.

5.2 Conclusions

Regarding our initial **Research Objective**, this study develops a system that can accurate classify discrete emotional states. This is achieved by using only peripheral physiological signals, specifically blood volume pulse (BVP) and electrodermal activity (EDA), recorded via a consumer-grade wearable device in a naturalistic gaming context.

Regarding **RQ1**, our study shows how, by combining these minimally intrusive signals with automated frame-level emotion labeling via FaceReader, a robust supervised learning pipeline can be developed. This is supported by achieving 86.01% accuracy and a macro-averaged AUC of 0.98 using Random Forest with SMOTE-based class balancing.

Compared to prior work that often relies on multimodal setups with EEG, ECG, or lab-constrained elicitation protocols, our approach offers competitive performance while prioritizing ecological validity, portability, and scalability. The use of commercial video games as emotion elicitation tasks introduces realistic, interactive, and varied emotional experiences, while automated facial annotation improves the scalability of the model, while helping reduce reliance on subjective self-reports.

Our research shows that, in regards to **RQ2**, how applying the SHAP technique can give insights into the inner workings of how model performs, taking a important step towards model transparency. This technique help us identify the more relevant component for the model prediction which narrow the array of possibilities for detecting and solving biases. SHAP results also provide guidance into how to best implement the model into a real-time system that could process the collected signals remotely.

The proposed pipeline can make use of a more invasive, with greater risks to privacy-preservation, emotion recognition software to develop an accurate system that uses signals which could be obtained from future gaming devices that can be minimally invasive and more privacy-preserving. We consider this to be a crucial contribution towards **RQ3**. Moreover, this pipeline helps future researchers adapt or update the system by changing the labeling software to suit their needs which could also help with solving biases that could arise.

These findings support the deployment of wearable-based affective computing systems for applications in education, adaptive gaming, mental health monitoring, and user experience research. As seen in Sect. 4.3, although the model occupies a memory space that can be a limitation to be fully integrated on a microcontroller, a very low latency implies that remote wearable measurings can be used together with an edge device or cloud-based computing to achieve real-time emotion recognition, specially when processing the in batches from 3 samples up.

Future work will benchmark this integration with its possible optimizations against our measured latency and further explore integrating additional available signals (e.g., skin temperature, accelerometry) from the Empatica EmbracePlus, as well as employing sequential deep learning architectures such as LSTMs or temporal CNNs to capture dynamic emotional transitions. Further validation in diverse user populations and real-world settings will strengthen the generalization and practical utility of this approach.

6 Limitations

While this study demonstrates promising results using physiological signals from a wearable device in a naturalistic gaming context, some limitations should be acknowledged. The participant sample consisted of 25 university students of similar age and educational background, which may limit the generalizability of the results to broader populations with greater age, cultural, or socio-economic diversity. Although theoretically this should not affect the model capabilities, future work should evaluate the model's performance across different demographic groups to ensure robustness and applicability in real-world deployments.

Additionally, while Noldus FaceReader provided automated, frame-level emotion labels that reduced reliance on self-reports. Although widely acknowledged as a reliable software by several studies [25, 38], its assessments of emotional states are taken as the initial labels of the system and thus the system could inherit possible biases. Nevertheless, the system pipeline is streamlined enough so that it can be easily updated with a future version of the Noldus software, or, if time is not of the essence, could be double-checked by an expert panel. Although initially discussed the necessity of manually verifying FaceReader outcomes, an initial comparison between our perceptions and those of FaceReader, together with the extensive literature supporting Noldus' system, provides grounds for trusting it output.

Future studies could strengthen annotation reliability by incorporating multiple independent reviewers and quantifying agreement, while a culturally diverse pool of participants would further test the system for possible biases.

Acknowledgements This publication is part of the R\&D\&I project PID2022-141172OA-I00, funded by MCIN/AEI /10.13039/501100011033/ and, where appropriate, by the European Regional Development Fund (ERDF) — "A way of making Europe". The authors are members of the research group "TEP108—Robotics and Computer Technology" at the University of Seville (Spain).

Autor contribution **Francisco Luna Perejón** contributed to conceptualization (supporting), investigation (equal), methodology (equal), resources (supporting), validation (equal), visualization (supporting), and writing (original draft and review & editing) **Miguel Civit** contributed to conceptualization (supporting), formal analysis (supporting), methodology (supporting), validation (supporting), and writing (original draft and review & editing). **Luis Muñoz-Saavedra** contributed equally to conceptualization, data curation, methodology, resources, validation and writing (original draft and review & editing). **Javier Civit-Masot** contributed equally to data curation, methodology, resources, software development, supported visualization and writing (original draft and review & editing). **Manuel Domínguez-Morales** contributed equally to formal analysis and resources, and supported investigation, methodology, and visualization. **Lourdes Miró-Amarante** contributed equally to conceptualization, investigation, methodology, visualization, and writing (original draft and review & editing), and led funding acquisition and supervision.

Funding Funding for open access publishing: Universidad de Sevilla/CBUA

Data availability The dataset used in this study is publicly available as the PaGER-Sync ADICVIDEO: Player Affective Gaming Experience & Responses - Synchronized ADICVIDEO Dataset. It has been formally described and published as a Data in Brief article. The dataset includes time-synchronized physiological signals (EDA, BVP, skin temperature) collected with an Empatica E4 wristband during naturalistic gameplay sessions, along with frame-level facial emotion annotations generated via FaceReader. It also contains pre-session psychometric questionnaire scores and demographic metadata. To protect participant privacy, the original webcam video recordings used for FaceReader analysis are not included in the published dataset. Only the resulting anonymized emotion labels are shared. All data are anonymized and organized in structured spreadsheet files to facilitate multimodal analysis. The dataset is accessible under a CC BY-NC-ND 4.0 license via the institutional repository idUS at: ●https://doi.org/10.12795/11441/174588. ● Handle: https://hdl.handle.net/11441/174588. Researchers are encouraged to consult the accompanying Data in Brief article for detailed documentation of the experimental design, data structure, and potential use cases.

Declarations

Conflict of interest There are no any conflicts of interest from authors.

References

1. Organization WH et al (2021) Comprehensive mental health action plan 2013–2030. Indian Journal of Psychiatry. https://doi.org/10.4103/indianjpsychiatry.indianjpsychiatry
2. Picard RW (1997) Affective computing. Pattern Analysis and Machine Intelligence. IEEE Transactions on 0001MIT Press. https://doi.org/10.7551/mitpress/1140.001
3. Wagh K, Vasanth K, Shinde S, Wadhwa L, Thakur A (2024) Emotion identification using eeg signal with reduced electrodes and time frequency parameters. International Journal of Performability Engineering 20(2):99–111. https://doi.org/10.23940/ijpe.24.02.p5.99111
4. Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39(6):1161–1178. https://doi.org/10.1037/h0077714
5. Kreibig SD (2010) Autonomic nervous system activity in emotion: A review. Biol Psychol 84(3):394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010
6. Boucsein W (2012). Electrodermal activity. https://doi.org/10.1007/978-1-4614-1126-0

7. Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K (2018) Introducing wesad, a multimodal dataset for wearable stress and affect detection. In: Proc. ICMI, pp 400–408 https://doi.org/10.1145/3242969.3242985
8. Li S, Deng W, Du J (2017) Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2852–2861. https://doi.org/10.1109/CVPR.2017.277
9. Jahangir R, Teh YW, Hanif F, Mujtaba G (2021) Deep learning approaches for speech emotion recognition: state of the art and research challenges. Multimedia Tools and Applications 80(16):23745–23812. https://doi.org/10.1007/s11042-020-09874-7
10. Koelstra S, Muhl C, Soleymani M, Lee J-S, Yazdani A, Ebrahimi T, Patras I (2011) Deap: A database for emotion analysis using physiological signals. IEEE Trans Affect Comput 3(1):18–31. https://doi.org/10.1109/T-AFFC.2011.15
11. Katsigiannis S, Ramzan N (2018) Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. IEEE J Biomed Health Inform 22(1):98–107. https://doi.org/10.1109/JBHI.2017.2688239
12. Eren E, Navruz TS (2022) Stress detection with deep learning using bvp and eda signals. 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). IEEE, pp 1–7 (**HORA55278.2022.9799933**)
13. Pidgeon M, Kanwal N, Murray N, Asghar M (2022) End-to-end emotion recognition using peripheral physiological signals. In: 35th International BCS Human-Computer Interaction Conference, BCS Learning & Development, pp 1–10 https://doi.org/10.14236/ewic/HCI2022.19
14. Domínguez-Jiménez JA, Campo-Landines KC, Martínez-Santos JC, Delahoz EJ, Contreras-Ortiz SH (2020) A machine learning model for emotion recognition from physiological signals. Biomed Signal Process Control 55:101646. https://doi.org/10.1016/j.bspc.2019.101646
15. Schmidt P, Reiss A, Dürichen R, Van Laerhoven K (2019) Wearable-based affect recognition-a review. Sensors 19(19):4079. https://doi.org/10.3390/s19194079
16. Qing C, Qiao R, Xu X, Cheng Y (2019) Interpretable emotion recognition using EEG signals. IEEE Access 7:94160–94170. https://doi.org/10.1109/ACCESS.2019.2928691
17. Muñoz-Saavedra L, Escobar-Linero E, Miró-Amarante L, Bohórquez MR, Domínguez-Morales M (2023) Designing and evaluating a wearable device for affective state level classification using machine learning techniques. Expert Syst Appl 219:119577. https://doi.org/10.1016/j.eswa.2023.119577
18. Indrasiri PL, Kankanhalli MS, Sadhya V (2024) VR Based Emotion Recognition Using Deep Multimodal Fusion With Biosignals Across Multiple Anatomical Domains. arXiv: 2412.02283
19. Bao G, Zhang J, Xu J, Li C (2021) Data augmentation model for eeg-based emotion recognition using a generative adversarial network. Front Comput Neurosci 15:723843. https://doi.org/10.3389/fncom.2021.723843
20. Liu J, Meng H, Nandi A, Li M (2016) Emotion detection from eeg recordings. In: Proc. 2016 12th Int. Conf. Nat. Comput, Fuzzy Syst. Knowl. Discovery (ICNC-FSKD), pp 1722–1727 https://doi.org/10.1109/FSKD.2016.7603437
21. D'Mello SK, Booth BM (2024) Affect Detection from Wearables in the "Real" Wild: Fact, Fantasy, or Somewhere In between? Available online. https://doi.org/10.1109/MIS.2022.3221854
22. Al-Nafjan A, Hosny M, Al-Ohali Y, Al-Wabil A (2017) Review and classification of emotion recognition based on eeg brain–computer interface system research: A systematic review. Appl Sci 7(12):1239. https://doi.org/10.3390/app7121239
23. Civit-Masot J, Luna-Perejón F, Muñoz-Saavedra L, Civit-Masot M, Dominguez-Morales M, Miró-Amarante L (2025) Pager-sync adicvideo: Player affective gaming experience & responses-synchronized adicvideo dataset. Data Brief 112115. https://doi.org/10.1016/j.dib.2025.112115
24. Empatica Inc (2025) EmbracePlus. https://www.empatica.com/research/e4/. Accessed: 2025-06-05
25. Skiendziel T, Rosch A, Schultheiss OC (2019) Assessing the convergent validity between the automated emotion recognition software Noldus FaceReader 7 and facial action coding system scoring. PLoS One. https://doi.org/10.1371/journal.pone.0223905
26. Jolliffe IT (2002). Principal Component Analysis. https://doi.org/10.1007/b98835
27. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemom Intell Lab Syst 2(1–3):37–52. https://doi.org/10.1016/0169-7439(87)80084-9
28. Abdi H, Williams LJ (2010) Principal component analysis. Wiley interdisciplinary reviews computational statistics 2(4):433–459. https://doi.org/10.1002/wics.101
29. Appelhans BM, Luecken LJ (2006) Heart rate variability as an index of regulated emotional responding. Rev Gen Psychol 10(3):229–240. https://doi.org/10.1037/1089-2680.10.3.229
30. Herborn KA, Graves JL, Jerem P, Evans NP, Nager R, McCafferty DJ, McKeegan DE (2015) Skin temperature reveals the intensity of acute stress. Physiology & Behavior 152:225–230. https://doi.org/10.1016/j.physbeh.2015.09.032
31. Ioannou S, Gallese V, Merla A (2014) Thermal infrared imaging in psychophysiology: Potentialities and limits. Psychophysiology 51(10):951–963. https://doi.org/10.1111/psyp.12243
32. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953
33. Goodfellow I, Bengio Y, Courville A (2016) Deep Learning, http://www.deeplearningbook.org

34. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 31, pp 3149–3157. Curran Associates Inc., Red Hook, NY, USA https://doi.org/10.5555/3294996.3295074
35. Hosmer DW Jr, Lemeshow S, Sturdivant RX (2013). Applied Logistic Regression. https://doi.org/10.1002/9781118548387
36. Kiranyaz S, Avci O, Abdeljaber O, Ince T, Gabbouj M, Inman DJ (2021) 1d convolutional neural networks and applications: A survey. Mech Syst Signal Process 151:107398. https://doi.org/10.1016/j.ymssp.2020.107398
37. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol 30, pp 4768–4777. https://doi.org/10.5555/3295222.3295230
38. Lewinski P, Den Uyl TM, Butler C (2014) Automated facial coding: validation of basic emotions and facs aus in facereader. J Neurosci Psychol Econ 7(4):227. https://doi.org/10.1037/npe0000028

Authors and Affiliations

Francisco Luna-Perejón[1] · Miguel Civit[2] · Luis Muñoz-Saavedra[1] · Javier Civit-Masot[1] · Manuel Domínguez-Morales[1] · Lourdes Miró-Amarante[1]

✉ Francisco Luna-Perejón
fluna1@us.es

1 Arquitectura y Tecnología de Computadores (ATC), E.T.S.I. Informática – Escuela Politécnica Superior, Universidad de Sevilla, Av. Reina Mercedes s/n, 41012 Seville, Spain

2 Departamento de Comunicación y Artes, Universidad Loyola Andalucía, Av. de las Universidades, 41704 Seville, Spain