Article

Not peer-reviewed version

Determining the Optimal Window Duration to Enhance Emotion Recognition Based on Galvanic Skin Response and Photoplethysmography Signals

Marcos F. Bamonte *, Marcelo Risk, Victor Herrero

Posted Date: 1 July 2024

doi: 10.20944/preprints202407.0058.v1

Keywords: Emotion Recognition; Galvanic Skin Response; Photoplethysmography; Optimal window duration; Nonlinear features; Machine Learning

Article

Determining the Optimal Window Duration to Enhance Emotion Recognition Based on Galvanic Skin Response and Photoplethysmography Signals

Marcos F. Bamonte [1]*, Marcelo Risk [2] and Victor Herrero [1]

[1] Facultad de Ingeniería - LIDTUA - CIC, Universidad Austral, Mariano Acosta 1611, B1629WWA, Pilar, Buenos Aires, Argentina; vherrero@austral.edu.ar

[2] Instituto de Medicina Traslacional e Ingeniería Biomédica (IMTIB) CONICET - IUHI - HIBA, Potosí 4240, C1199ACL, Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina; marcelo.risk@hospitalitaliano.org.ar

* Correspondence: mbamonte@austral.edu.ar

Abstract: Automatic emotion recognition using portable sensors is gaining attention due to its potential use in real-life scenarios. Existing studies have not explored Galvanic Skin Response and Photoplethysmography sensors exclusively for emotion recognition using nonlinear features with machine learning (ML) classifiers such as Random Forest, Support Vector Machine, Gradient Boosting Machine, K-Nearest Neighbor, and Decision Tree. In this study, we proposed a genuine window sensitivity analysis on a continuous annotation dataset to determine the window duration and percentage of overlap that optimize the classification performance using ML algorithms and nonlinear features, namely Lyapunov Exponent, Approximate Entropy, and Poincaré's indices. We found an optimum window duration of 11 seconds and 16 seconds for valence and arousal, respectively, with 50% overlap, and achieved an accuracy of 0.76 in both dimensions. In addition, we proposed a Strong Labeling Scheme that kept only the extreme values of the labels, which raised the accuracy score to 0.92. Under certain conditions mentioned, traditional ML models offer a good compromise between performance and low computational cost. Our results suggest that well-known ML algorithms can still contribute to the field of emotion recognition, provided that window duration, overlap percentage, and nonlinear features are carefully selected.

Keywords: Emotion Recognition; Galvanic Skin Response; Photoplethysmography; Optimal window duration; Nonlinear features; Machine Learning

1. Introduction

With the growing proliferation of wearable sensors capable of uploading biosignal data to the cloud, automatic emotion recognition has acquired significant interest due to its potential applications in education, psychology, well-being, medicine, neuroscience, driver safety, and other fields [1–5].

The most common biosensors are electrocardiography (ECG), respiration (RESP), electroencephalography (EEG), galvanic skin response (GSR) or electrodermal activity (EDA), electrooculography (EOG), photoplethysmography (PPG) or blood volume pulse (BVP), electromyography (EMG), and skin temperature (SKT or TEMP) [2,3]. Not all of them are comfortable, user-friendly, or portable, which makes them ill-suited to be employed outside a laboratory environment, at least with the current technological development.

Among the mentioned biosensors, Galvanic Skin Response (GSR) and Photoplethysmography (PPG) stand out as portable, non-invasive sensors capable of gathering larger volumes of data over time due to their ease of use. Although there are not many portable GSR and PPG sensors capable of collecting clinical-quality data currently, sensors with improved signal quality are expected to emerge in the future [6]. Thus, in this study, we focus on automatic emotion recognition employing only GSR and PPG biosignals. GSR sensors typically measure skin electrical conductance using two electrodes, usually placed on the fingers. Skin conductance is linked to sweating, which in turn is connected to emotions [7,8]. On the other hand, PPG sensors indirectly measure heart rate and other associated metrics, which are also linked to emotions [7,9]. They are typically worn on the wrist.

 doi:10.20944/preprints202407.0058.v1

Emotion recognition is carried out by applying machine learning algorithms directly to the biosignals or some set of extracted features when an individual is subjected to some affect elicitation stimulus (e.g., video clips, images) [3]. The individual usually annotates emotions on two continuous scales, i.e., valence and arousal, typically ranging from 1 to 9. Valence denotes how pleasant or unpleasant the emotion is, while arousal represents the intensity [3]. Valence and arousal are usually treated as two independent classification problems [10].

Additionally, data annotation can be discrete or continuous [11–14]. In the former case, labels are recorded in an indirect, post-hoc manner, e.g., one label is annotated after a video clip of 60 seconds is shown. In the latter case, data labels are annotated with higher frequencies. Most publicly available datasets follow a discrete annotation paradigm.

The process of emotion elicitation and labeling usually takes no less than an hour per individual, including participant instruction, trials to familiarize with the system, baseline recordings, stimulus presentation, and annotations. This induces fatigue in participants. As a result, datasets typically do not have many samples per participant. This is a significant problem in the emotion recognition field, but it can be addressed with proper segmentation of the labels, at least until larger datasets become available.

Regarding the selection of features to extract from biosignals, there is no consensus on the optimum set that maximizes the accuracy of emotion recognition in every situation. The selection of features is typically problem-dependent [3]. Nonetheless, features from temporal, statistical, and nonlinear domains extracted from GSR and PPG signals have yielded very good results [1,15,16]. A particular challenge is finding the set that combines GSR and PPG extracted features to yield optimal performance.

Some existing works on emotion recognition are based solely on GSR and PPG. Martínez et al. [17] proposed a stack of two Convolutional Neural Networks (CNNs) followed by a simple perceptron to recognize discrete emotions (relaxation, anxiety, excitement, and fun) using GSR and PPG modalities from participants playing a predator/prey game. They found that the proposed deep learning model, which automatically and directly extracts features from the raw data, outperforms models utilizing known statistically ad-hoc extracted features, attaining an accuracy of 0.75. Ayata et al. [18] proposed a music recommendation system based on emotions using the DEAP dataset [11], which utilizes music videos to elicit emotions. They achieved accuracies of 0.72 and 0.71 on arousal and valence, respectively, by feeding a Random Forest (RF) classifier with statistical features. The work studied the effect of window duration size for GSR and PPG separately and found that 3-second windows performed better for GSR, while 8-second windows performed better for PPG. Kang et al. [19] presented a signals-based labeling method that involved windowing (data were window-sliced in 1-pulse units) and observer-annotated data. They applied a 1D convolutional neural network to recognize emotions and obtained accuracies of 79.18% and 74.84% on the MERTI-Apps dataset [20] for arousal and valence, respectively, while achieving 81.33% and 80.25% on arousal and valence, respectively, using the DEAP dataset.

Goshvarpour et al. [16] implemented a Probabilistic Neural Network (PNN) to recognize emotions based on nonlinear features. Approximate Entropy, Lyapunov Exponent, and Poincaré indices (PI) were extracted and fed to the PNN. They validated the experiment using the DEAP dataset and obtained 88.57% and 86.8% for arousal and valence, respectively. Domínguez-Jiménez et al. [21] conducted an experiment with 37 volunteers employing wearable devices. They were able to recognize three emotions (i.e., amusement, sadness, and neutral) with an accuracy of 100% when a linear Support Vector Machine (SVM) classifier was trained with statistical features selected by feature selection methods such as Genetic Algorithm (GA) or Random Forest Recursive Feature Selection (RF-RFE). In addition, in our previous work [22], we adopted a robust labeling scheme that discarded neutral values of valence and arousal, keeping only extreme values (i.e., ranges from [1-3] and [7-9]). As we were interested in testing the generalization skills of certain algorithms, such as SVM, K-Nearest Neighbor (KNN), and Gradient Boosting Machine (GBM), in a subject-dependent emotion classification context,

 doi:10.20944/preprints202407.0058.v1

we tested the same model parameters on two datasets: DEAP and K-EmoCon [13]. The former employs non-portable sensors, while the latter uses wearable sensors. We found that accuracies of 0.7 and an F1-score of 0.57 are attainable, but this comes at the expense of discarding some samples.

In our study, we employed a continuous annotation dataset. The purpose of this decision is twofold: to use a greater number of samples in the model's training and to perform genuine window segmentation on the data and annotations, allowing for better emotion capture with different window duration sizes. We carried out a sensitivity study on window duration size and percentage overlap to find the optimal values that result in better recognition performance. Additionally, we compared the performance of temporal-statistical and nonlinear features. Moreover, different labeling schemes were adopted to explore how accuracy increased as the thresholds for label binarization were raised.

Finally, the main aim of our work was to find the window duration size and percentage of overlap that optimized emotion recognition performance using only GSR and PPG while employing low computational cost algorithms. We found that this can be accomplished provided data annotation is continuous and nonlinear features are used.

2. Materials and Methods

To conduct this study we employed the CASE dataset [14], which features continuous annotations sampled at 20 Hz. The annotations were made using a joystick-based interface to rate each stimulus video on the valence and arousal continuous scale, ranging from 1 to 9. GSR and PPG raw data were sampled at 1000 Hz.

Each participant was subjected to 8 film clips, each lasting 1.5 to 3.5 minutes. During the stimulus, the participant annotated the video on a 2D plane in real-time. The axes of the plane are valence and arousal. A total of 30 participants conducted the experiment, 15 male and 15 female.

2.1. Data Preprocessing

We made use of linear interpolated data and annotations to mitigate data latency issues that occurred during data acquisition, ensuring they were effectively sampled at regular intervals of 1 ms and 50 ms, respectively [14]. Subsequently, we applied a baseline mean-centered normalization as suggested by [15], following this equation:

$$S_n = S - mean(S_r) \tag{1}$$

where S_r represents signals acquired during the last 60 seconds of the start-video [14].

After normalization, we applied a fourth-order Butterworth low-pass filter with a cut-off frequency of 3 Hz to the GSR signal and a third-order Butterworth band-pass filter with cutoff frequencies of 0.5-8 Hz to the PPG signal, using the Neurokit package [23].

To test the effects of window duration and percentage of overlap, we segmented the data and continuous labels into various window sizes ($W_{size} \in [1, 3, 5, 7, 9...29]$ seconds) and overlap percentage ($O_{lap} \in [0, 25, 50]\%$). Before segmentation, annotations were upsampled to match the sample rate of the PPG and GSR signals.

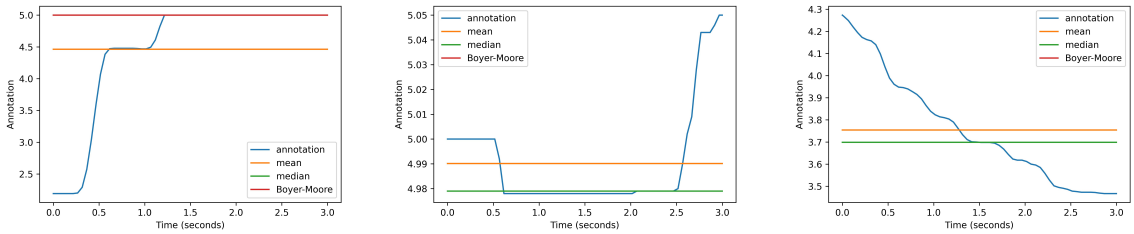
2.2. Labels Mapping

Because data labels are continuous and can vary within a given segment, each annotation segment should be replaced by a particular value, to train different models and make classification possible. We replaced each segment with its median value. We think this method better represents user labeling than replacing the annotation segment with its mean value, or even the majority value given by the Boyer-Moore voting algorithm [24].

When there's a clear majority in favor of a particular label, the median value provides a result close to the longest horizontal line. In such cases, both the Boyer-Moore and median methods yield the same or almost the same result, while the mean may produce a different value (see Figure 1a). On the other hand, when there's no clear majority, the Boyer-Moore method doesn't yield a result, and the

 doi:10.20944/preprints202407.0058.v1

mean sometimes deviates significantly from the most selected labels, whereas the median tends to be closer to the labels corresponding to the longest horizontal subsegments, better representing the selected majority of labels for the given segment (see Figure 1b). In the extreme situation where there is no label with more than one vote, the median and the mean yield approximately the same results (see Figure 1c)



(a) (b) (c)

Figure 1. Segment Labels mapping: from continuous multivalued labels to one value annotation per segment. **a** Clear majority. **b** No clear majority. **c** No clear majority: extreme case

2.3. Labeling Schemes

Once the continuous labels in a segment were replaced with its median value, a slight variant of the bipartition labeling schemes (BLS) mentioned by [25] is adopted. Labels were binarized according to three different schemes: classic, weak, and strong. Each scheme is detailed in Table 1.

Table 1. Labeling schemes.

Classic labeling scheme (CLS)	Weak labeling scheme (WLS)	Strong labeling scheme (SLS)
$\forall x : x \leq 5 \Rightarrow \text{map } x \rightarrow 0$	$\forall x : x \leq 4 \Rightarrow \text{map } x \rightarrow 0$	$\forall x : x \leq 3 \Rightarrow \text{map } x \rightarrow 0$
$\forall x : x > 5 \Rightarrow \text{map } x \rightarrow 1$	$\forall x : x \geq 6 \Rightarrow \text{map } x \rightarrow 1$	$\forall x : x \geq 7 \Rightarrow \text{map } x \rightarrow 1$
	discard $x : 4 < x < 6$	discard $x : 3 < x < 7$

2.4. Data Splitting

To handle imbalanced data, all the data and its corresponding labels were split into train and test datasets, following a stratified K-Fold cross-validation strategy with 10 folds. We used 20% of the data for testing.

Finally, signals from the GSR and PPG sensors were standardized following [15]:

$$S_d = \frac{S_n - mean(S_n)}{\sigma\{S_n\}} \tag{2}$$

where $\sigma\{S_n\}$ is the standard deviation of the baseline mean-centered normalized signal. Finally, standardization was fitted on the training dataset and applied to the training and testing dataset.

2.5. Feature Extraction

Based on the good results and methodology described in [16], we extracted several nonlinear features for GSR and PPG, as these signals exhibit chaotic and nonlinear behaviors. Specifically, we extracted Approximate Entropy (ApEn), Lyapunov Exponent (LE), and some Poincaré indices (PI). While ApEn measures the complexity or irregularity of a signal [16], the LE of a time series indicates whether the signal presents chaotic behavior or, conversely, has a stable attractor. On the other hand, specific indices from Lagged Poincaré Plots quantify the PPG and GSR attractors. The extracted features can be seen in Table 2.

 doi:10.20944/preprints202407.0058.v1

Table 2. Extracted nonlinear features from GSR and PPG.

Parameter	Description
LE	Lyapunov exponent (Rosenstein et al. method [26])
ApEn	Approximate entropy
$SD1_l$	Poincaré plot standard deviation perpendicular to the line of identity [9], for lag l[1,Goshvarpour2020]
$SD2_l$	Poincaré plot standard deviation along the identity line [9], for lag l[1]
$SD12_l$	Ratio of SD1-to-SD2, for lag l[1]
S_l	Area of ellipse described by SD1 and SD2, for lag l[1,Goshvarpour2020]

[1] Indices were computed for lags l of one and ten.

A total of 20 nonlinear features were extracted for each window segment: 10 for GSR and 10 for PPG. These features are LE, ApEn, $SD1_1$, $SD2_1$, $SD12_1$, S_1, $SD1_{10}$, $SD2_{10}$, $SD12_{10}$, and S_{10}.

To compare the ability to extract relevant information from the physiological signals between two different feature extraction domains, well-known temporal-statistical features were also extracted, as shown in Table 3.

Table 3. Extracted temporal-statistical features from GSR and PPG.

Parameter	Description
GSR,Godin2015:	
Avg_d	Average of the derivative
Neg_s	% of neg. samples in the derivative
L_m	number of local minima
PPG,Godin2015[28]:	
BPM	Beats per minute
IBI	Mean inter-beat interval
SDNN	Standard deviation of intervals between adjacent beats
RMSSD	Root mean square of successive differences between neighbouring heart beat intervals
SDSD	Standard deviation of successive differences between neighbouring heart beat intervals

A total of 8 temporal-statistical features were extracted for each window segment: 3 for GSR and 5 for PPG.

For comparison purposes, we also trained a Convolutional Neural Network with a Single Layer Perceptron Classifier (CNN-SLP) algorithm, as detailed in [22]. The CNN-SLP is a variation of the model proposed by [17]. It automatically extracts features directly from the GSR and PPG time series, following a representation learning paradigm [29]. These learned features from PPG and GSR are fused into a unified vector, which serves as the input for a Single Layer Perceptron classifier (See Figure 2).

 doi:10.20944/preprints202407.0058.v1

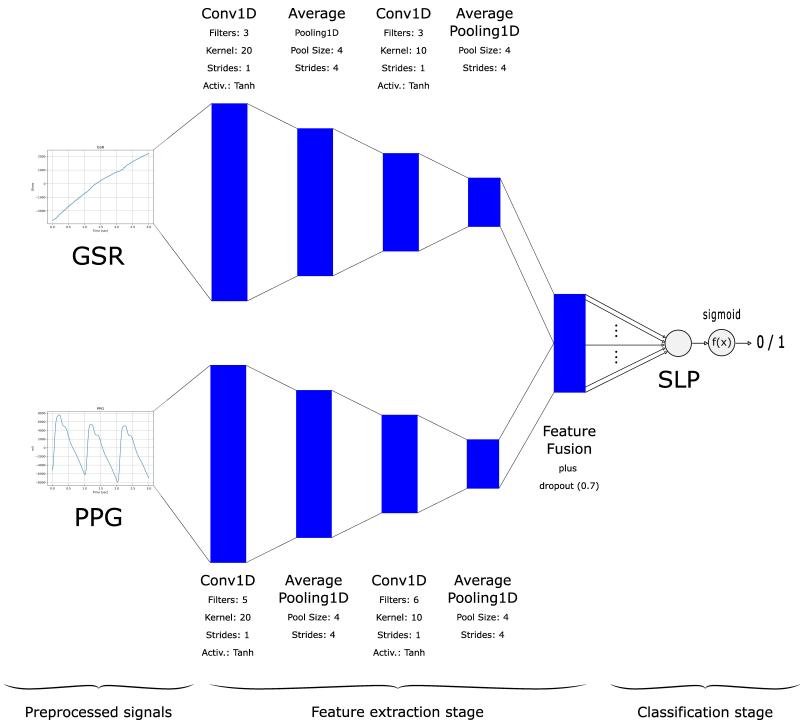


Figure 2. Convolutional Neural Network with a Single Layer Perceptron Classifier (CNN-SLP). First published in IFMBE Proceedings, Volume 1, pages 23-35, 2024 by Springer Nature [22].

2.6. Algorithm Training and Performance Evaluation

In this study, six algorithms were trained: Decision Tree (DT), KNN, RF, SVM, GBM, and CNN-SLP. Except for the latter, all other algorithms can be considered shallow learning models [30].

As we were interested in performing a subject-dependent emotion classification [3], and assessing the generalization ability of each algorithm, we employed the same hyperparameters utilized in [22], tested on the DEAP and K-Emocon datasets [11,13]. The set of chosen hyperparameters can be seen in Table 4.

Table 4. Algorithms hyperparameters.

Algorithm	Hyperparameters
KNN	neighbors = 5
DT	criterion = gini
RF	estimators = 5
SVM	regularization $C = 0.1$
GBM	estimators = 5
CNN-SLP	See Figure 2

To assess the performance, we employed the accuracy (ACC), the unweighted average recall (UAR) [31], and the F1-score (F1). Both UAR and F1 are well-suited for imbalanced datasets.

Because our approach is subject-dependent emotion recognition, we computed these metrics for each participant's test dataset and calculated the average across all participants, as shown in the next section.

2.7. Code Availability

All the simulations conducted in this work were coded in Python and have been made freely available at https://github.com/mbamonteAustral/Emotion-Recognition-Based-on-Galvanic-Skin-Response-and-Photoplethysmography-Signals.git.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 July 2024 doi:10.20944/preprints202407.0058.v1

3. Results

Table 5 shows the best performance obtained for all trained algorithms. Although metrics are computed for a particular window size and overlap, RF outperformed all other algorithms in nearly every simulation we ran for this work, in terms of both valence and arousal. For this reason, we will focus our results mainly on this algorithm. An accuracy of 0.76 was attained for the two aforementioned dimensions.

Because we worked with imbalanced data, we added a dummy classifier as a baseline, which makes predictions based on the most frequent class (i.e., always returns the most frequent class in the observed labels). This baseline helps to appreciate the skill of the different tested models.

Table 5. Mean accuracy (ACC), unweighted average recall (UAR), and F1-score (F1) for binary affect classification using GSR and PPG with a classical labeling scheme (CLS), optimal window size, and overlap.

	Valence [1]			Arousal [2]		
Classifier	UAR	ACC	F1	UAR	ACC	F1
KNN	0.70	0.73	0.69	0.72	0.74	0.72
DT	0.70	0.73	0.70	0.70	0.72	0.70
RF	**0.73**	**0.76**	**0.72**	**0.74**	**0.76**	**0.73**
SVM	0.57	0.67	0.50	0.61	0.68	0.56
GBM	0.68	0.74	0.66	0.68	0.74	0.68
CNN-SLP	0.57	0.62	0.57	0.61	0.65	0.60
Baseline	0.50	0.61	0.37	0.5	0.59	0.37

[1] $W_{size} = 16$ sec, $O_{lap} = 50\%$
[2] $W_{size} = 11$ sec, $O_{lap} = 50\%$

3.1. Impact of Window Duration Size and overlap

As can be seen in Figure 3, accuracy decreases with increasing window size, both in valence and arousal, when there's no overlap. A similar trend is observed in the accuracy when the $O_{lap} = 25\%$. However, when the overlap is 50%, accuracy decreases more slowly. It can be shown that accuracy decreases for window duration sizes greater than 30 seconds. For $O_{lap} = 50\%$, there are optimum window duration sizes of 11 and 16 seconds for arousal and valence, respectively, in terms of ACC, UAR, and F1 scores. Moreover, from our results, all metric scores increase as the overlap increases.

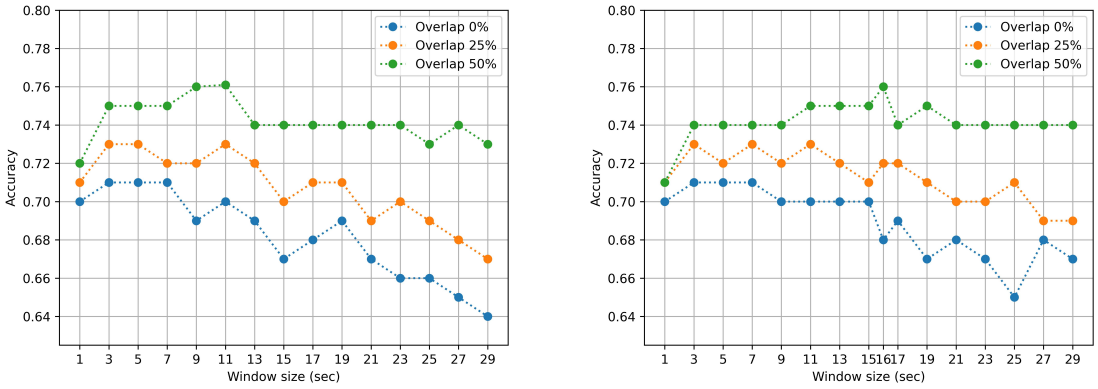
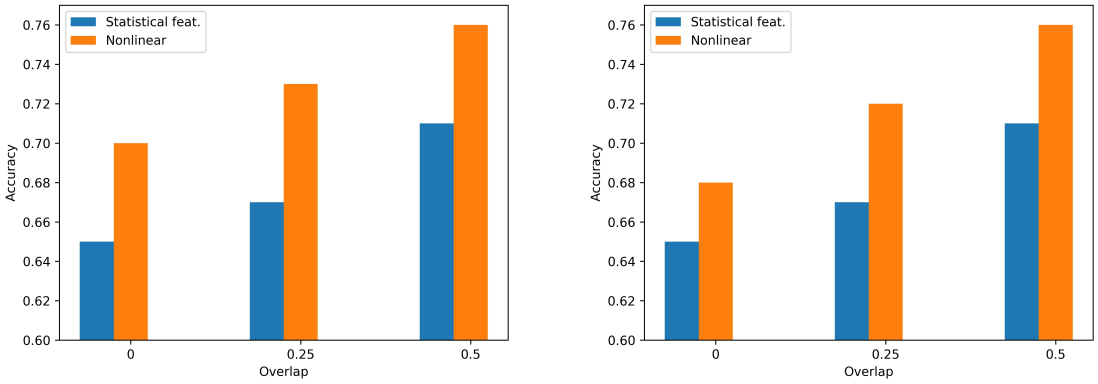


Figure 3. Accuracy performance for different window duration sizes and percentages of overlap, employing Random Forest (RF). **a** Arousal. **b** Valence.

 doi:10.20944/preprints202407.0058.v1

3.2. Features Domain Performance Comparison

Both temporal-statistical and nonlinear features were extracted for different window duration sizes and percentages of overlap. In every case, nonlinear features outperformed temporal-statistical features in terms of performance. Figure 4 illustrates this comparison. In the best case, nonlinear features yielded an accuracy of 0.76, while temporal-statistical features achieved 0.71.

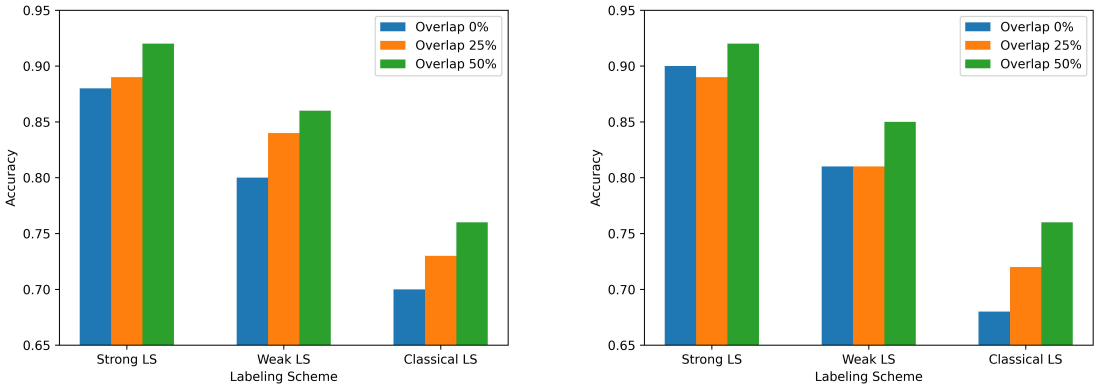


(a) (b)

Figure 4. Features domain performance comparison for optimum window size, employing Random Forest. **a** Arousal ($W_{size} = 11$). **b** Valence ($W_{size} = 16$).

3.3. Labeling Schemes comparison

The Strong Labeling Scheme proved to be more accurate than the WLS and the CLS. The best accuracy of 0.92 for valence and arousal was obtained employing the SLS and $O_{lap} = 50\%$, as can be seen in Figure 5.



(a) (b)

Figure 5. Labeling scheme performance comparison for optimum window size, employing Random Forest. **a** Arousal ($W_{size} = 11$). **b** Valence ($W_{size} = 16$).

4. Discussion

The main finding is the identification of different optimal window durations for arousal and valence, which are 11 seconds and 16 seconds, respectively, for the CASE dataset, using 50% overlap. It is worth mentioning that these optimal values arise from the fusion of PPG and GSR.

As the overlap increases, four different effects can be identified: (a) accuracy improves, (b) the optimal window duration value rises, (c) an approximate plateau appears for window durations

 doi:10.20944/preprints202407.0058.v1

greater than the optimum in the 50% overlap situation, and (d) performance decreases slowly over longer window durations (the curve flattens). It is important to note that accuracy decreases as the window duration grows when there is no overlap, consistent with other studies [18,32].

These effects could be due to the reinforcement in the model's training caused by the overlap, which allows capturing emotional information that would otherwise be missed in the no-overlap situation with the same window duration. Interestingly, accuracy doesn't fall abruptly as window durations increase.

Ayata et al. [18] reported maximum accuracy for one-second overlapped windows of $W_{size} = 3$ seconds and $W_{size} = 8$ seconds, respectively, using only GSR or PPG. They extracted temporal and statistical features to train the models. The optimal window duration held for both valence and arousal.

On the other hand, Zhang et al. [32] achieved maximum accuracy for a non-overlapped window of $W_{size} = 2$ seconds using ECG, PPG, GSR, and Heart Rate (HR) on the CASE dataset for both valence and arousal. For the MERCA dataset, non-overlapped windows of $W_{size} = 2$ seconds and $W_{size} = 4$ seconds achieved maximum accuracy for arousal and valence, respectively.

The difference in our results might be due to the fusion of PPG and GSR features, the use of nonlinear features, and the continuous annotations binarized using the median method. This suggests that the optimal window duration depends on the biosignals and the particular processing pipeline employed to train the models (i.e., preprocessing method, set of extracted features, continuous vs. discrete annotations, label mapping method), but especially on the percentage of overlap.

Regarding extracted features, nonlinear outperformed temporal-statistical features in every situation, suggesting a greater skill in extracting emotional information from the biosignals. This is consistent with current trends in nonlinear feature extraction [1,16].

Although [16] showed better accuracy scores using nonlinear feature fusion for GSR and PPG and a PNN as the classifier, we found that shallow learning algorithms offer a good compromise between performance and low computational cost.

Concerning the labeling scheme, SLS and WLS performed better than CLS, as expected, because CLS only kept extreme values of the labels while discarding values around the neutral. This facilitated pattern recognition and model training but came at the expense of discarding samples.

Finally, we achieved better results than in our previous work [22], although we applied the same algorithms configured with the same hyperparameter configuration. This suggests that genuine windowing performed on a continuous annotations dataset, combined with the extraction of nonlinear features, proved to extract more emotional information from the biosignals than using a discrete annotation dataset with temporal-statistical features. A comparison with related studies can be seen in Table 6.

4.1. Limitations

It is worth mentioning that we employed a specific set of algorithm hyperparameters and nonlinear features on a particular dataset. Other combinations should be tested on several continuous annotation datasets to determine which combination exhibits better generalization skills. Additionally, different processing pipelines might yield varying optimum window durations, as the optimal value can depend on the overall processing method.

Some nonlinear features are computationally more costly than others (e.g., the Lyapunov Exponent takes longer to compute than Poincare indices and Approximate Entropy). In future work, we plan to explore other nonlinear features to optimize computational efficiency.

We employed the CASE dataset (see Section 1) to test the working hypotheses on a continuous annotation dataset, allowing for genuine window segmentation. Although this dataset uses FDA-approved sensors, some of its instruments (e.g., the ADC module) are not particularly suited for real-life scenarios. We considered two recent datasets, namely Emognition [33] and G-REx [34], which use PPG and GSR wearable devices, but their annotation method is not continuous.

 doi:10.20944/preprints202407.0058.v1

Table 6. Comparison with other other related studies.

Author	Modalities	Windowing / Overlap[1]	Features[2]	Classifier	Computational Cost	ACC[3]
Goshvarpour et al. [16]	PPG, GSR	-	NL	PNN	High	A: 88.5% V: 86.8%
Martínez et al. [17]	PPG, GSR	W: Yes O: No	A	SCAE[4]	High	<75.0%[5]
Ayata et al. [18]	PPG, GSR	W: Yes O: Yes	ST	RF	Low	A: 72.0% V: 71.0%
Kang et al. [19]	PPG, GSR	W: Yes O: No	A	CNN	High	A: 81.3% V: 80.2%
Domínguez-Jiménez et al. [21]	PPG, GSR	W: Yes O: No	ST, NL[6]	SVM	Low	100%[7]
Our previous work [22]	PPG, GSR	W: No O: No	TST	SVM	Low	A: 73.0% V: 71.0%[8]
Zitouni et al. [15]	PPG, GSR and HR	W: Yes O: Yes	ST, NL	LSTM	High	A: 92.0% V: 95.0%
Santamaría-Granados et al. [35]	ECG-GSR	W: Yes O: Yes	TST, F, NL	DCNN	High	A: 76.0% V: 75.0%
Cittadini et al. [36]	GSR, ECG, RESP	W: Yes O: No	TST	KNN	Low	A: 78.5% V: 80.5%
Zhang et al. [32]	PPG, GSR, ECG, HR	W: Yes O: No	A	CorrNet[9]	High	A: 74.0% V: 76.3%
Present work	PPG, GSR	W: No O: Yes	NL	RF	Low	A: 76.0% V: 76.0%

[1] W: Windowing, O: Overlap
[2] NL: Nonlinear, A: Automatic learned, ST: Statistical, T: Temporal, TST: Temporal-statistical
[3] A: Arousal, V: Valence
[4] SCAE: Stacked Convolutional Auto-encoders
[5] Discrete emotions: Relaxation, Anxiety, Excitement, Fun
[6] Applied only to GSR signal
[7] Discrete emotions: Amusement, Sadness, and Neutral
[8] Employing Strong Labeling Scheme
[9] Deep Learning algorithm

As more commercial FDA-approved GSR and PPG wearable sensors continue to emerge, the availability of data samples for model training in real-life situations will increase [6][37]. This increased data availability might make continuous annotation unnecessary. In this situation, SLS could be employed when there is interest only in very definite emotions (high or low arousal, high or low

 doi:10.20944/preprints202407.0058.v1

valence, no neutral values). Samples discarded by this scheme might be compensated for by a larger data volume, keeping the model training robust.

5. Conclusions

In this work, we performed a genuine window sensitivity study to determine the window duration that optimized emotion recognition accuracy for valence and arousal, based only on PPG and GSR. We tested different percentages of overlap in a continuous annotation dataset. Additionally, we compared the performance of nonlinear and temporal-statistical features, verifying that the former extracts more emotional information from the biosignals.

We confirmed that recognizing emotions with acceptable accuracy is possible using only the mentioned biosignals, provided continuous labeling, nonlinear features, optimized window size, and overlap percentage are employed. Under these conditions, well-known shallow learning algorithms offer a good compromise between performance and low computational cost. Furthermore, if the SLS scheme is used, excellent performance can be achieved, although some samples may be discarded. This issue could be mitigated by employing wearables in longitudinal real-life scenarios, given their potential to gather larger volumes of data.

Author Contributions: Conceptualization, M.B, V.H and M.R.; methodology, M.B, V.H and M.R.; software, M.B.; validation, M.B., V.H. and M.R.; formal analysis, M.B.; investigation, M.B.; resources, V.H.; data curation, M.B.; writing—original draft preparation, M.B.; writing—review and editing, M.B, V.H. and M.R; visualization, M.B.; supervision, V.H, M.R.; project administration, V.H; funding acquisition, V.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no funding.

Data Availability Statement: The dataset used in this work is openly available in GitLab at https://gitlab.com/karan-shr/case_dataset. Follow Section 2.7 instructions to reproduce our simulations.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Khare, S.K.; Blanes-Vidal, V.; Nadimi, E.S.; Acharya, U.R. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Inf. Fusion* **2023**, *102*, 102019. doi:10.1016/j.inffus.2023.102019.
2. Dzedzickis, A.; Kaklauskas, A.; Bucinskas, V. Human emotion recognition: Review of sensors and methods. *Sensors (Switzerland)* **2020**, *20*. doi:10.3390/s20030592.
3. Bota, P.J.; Wang, C.; Fred, A.L.N.; Placido Da Silva, H. A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals. *IEEE Access* **2019**, *7*, 140990–141020. doi:10.1109/ACCESS.2019.2944001.
4. Schmidt, P.; Reiss, A.; Dürichen, R.; Laerhoven, K.V. Wearable-Based Affect Recognition—A Review. *Sensors* **2019**, *19*, 4079. doi:10.3390/s19194079.
5. Davoli, L.; Martalò, M.; Cilfone, A.; Belli, L.; Ferrari, G.; Presta, R.; Montanari, R.; Mengoni, M.; Giraldi, L.; Amparore, E.G.; Botta, M.; Drago, I.; Carbonara, G.; Castellano, A.; Plomp, J. On driver behavior recognition for increased safety: A roadmap, 2020. doi:10.3390/safety6040055.
6. Gomes, N.; Pato, M.; Lourenço, A.R.; Datia, N. A Survey on Wearable Sensors for Mental Health Monitoring. *Sensors (Basel).* **2023**, *23*. doi:10.3390/S23031330.
7. Kreibig, S.D. Autonomic nervous system activity in emotion: A review, 2010. doi:10.1016/j.biopsycho.2010.03.010.
8. van Dooren, M.; de Vries, J.J.J.; Janssen, J.H. Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiol. Behav.* **2012**, *106*, 298–304. doi:10.1016/j.physbeh.2012.01.020.
9. Rinella, S.; Massimino, S.; Fallica, P.G.; Giacobbe, A.; Donato, N.; Coco, M.; Neri, G.; Parenti, R.; Perciavalle, V.; Conoci, S. Emotion Recognition: Photoplethysmography and Electrocardiography in Comparison. *Biosensors* **2022**, *12*, 811. doi:10.3390/bios12100811.
10. Huang, Y.; Yang, J.; Liu, S.; Pan, J. Combining facial expressions and electroencephalography to enhance emotion recognition. *Futur. Internet* **2019**, *11*, 105. doi:10.3390/fi11050105.

11. Koelstra, S.; Mühl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. DEAP: A database for emotion analysis; Using physiological signals. *IEEE Trans. Affect. Comput.* **2012**, *3*, 18–31. doi:10.1109/T-AFFC.2011.15.
12. Miranda-Correa, J.A.; Abadi, M.K.; Sebe, N.; Patras, I. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Trans. Affect. Comput.* **2021**, *12*, 479–493, [1702.02510]. doi:10.1109/TAFFC.2018.2884461.
13. Park, C.Y.; Cha, N.; Kang, S.; Kim, A.; Khandoker, A.H.; Hadjileontiadis, L.; Oh, A.; Jeong, Y.; Lee, U. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Sci. Data* **2020**, *7*, 293. doi:10.1038/s41597-020-00630-y.
14. Sharma, K.; Castellini, C.; van den Broek, E.L.; Albu-Schaeffer, A.; Schwenker, F. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Sci. Data* **2019**, *6*, 196, [1812.02782]. doi:10.1038/s41597-019-0209-0.
15. Zitouni, M.S.; Park, C.Y.; Lee, U.; Hadjileontiadis, L.J.; Khandoker, A. LSTM-Modeling of Emotion Recognition Using Peripheral Physiological Signals in Naturalistic Conversations. *IEEE J. Biomed. Heal. Informatics* **2023**, *27*, 912–923. doi:10.1109/JBHI.2022.3225330.
16. Goshvarpour, A.; Goshvarpour, A. The potential of photoplethysmogram and galvanic skin response in emotion recognition using nonlinear features. *Phys. Eng. Sci. Med.* **2020**, *43*, 119–134. doi:10.1007/s13246-019-00825-7.
17. Martinez, H.P.; Bengio, Y.; Yannakakis, G. Learning deep physiological models of affect. *IEEE Comput. Intell. Mag.* **2013**, *8*, 20–33. doi:10.1109/MCI.2013.2247823.
18. Ayata, D.; Yaslan, Y.; Kamasak, M.E. Emotion Based Music Recommendation System Using Wearable Physiological Sensors. *IEEE Trans. Consum. Electron.* **2018**, *64*, 196–203. doi:10.1109/TCE.2018.2844736.
19. Kang, D.H.; Kim, D.H. 1D Convolutional Autoencoder-Based PPG and GSR Signals for Real-Time Emotion Classification. *IEEE Access* **2022**, *10*, 91332–91345. doi:10.1109/ACCESS.2022.3201342.
20. Maeng, J.H.; Kang, D.H.; Kim, D.H. Deep Learning Method for Selecting Effective Models and Feature Groups in Emotion Recognition Using an Asian Multimodal Database. *Electron. 2020, Vol. 9, Page 1988* **2020**, *9*, 1988. doi:10.3390/ELECTRONICS9121988.
21. Domínguez-Jiménez, J.A.; Campo-Landines, K.C.; Martínez-Santos, J.C.; Delahoz, E.J.; Contreras-Ortiz, S.H. A machine learning model for emotion recognition from physiological signals. *Biomed. Signal Process. Control* **2020**, *55*, 101646. doi:10.1016/j.bspc.2019.101646.
22. Bamonte, M.F.; Risk, M.; Herrero, V. Emotion Recognition Based on Galvanic Skin Response and Photoplethysmography Signals Using Artificial Intelligence Algorithms. Advances in Bioengineering and Clinical Engineering; Ballina, F.E.; Armentano, R.; Acevedo, R.C.; Meschino, G.J., Eds.; Springer Nature Switzerland: Cham, 2024; pp. 23–35. doi:10.1007/978-3-031-61960-1_3.
23. Makowski, D.; Pham, T.; Lau, Z.J.; Brammer, J.C.; Lespinasse, F.; Pham, H.; Schölzel, C.; Chen, S.H. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behav. Res. Methods* **2021**, *53*, 1689–1696. doi:https://doi.org/10.3758/s13428-020-01516-y.
24. Boyer, R.S.; Moore, J.S., MJRTY—A Fast Majority Vote Algorithm. In *Automated Reasoning: Essays in Honor of Woody Bledsoe*; Boyer, R.S., Ed.; Springer Netherlands: Dordrecht, 1991; pp. 105–117. doi:10.1007/978-94-011-3488-0_5.
25. Menezes, M.L.; Samara, A.; Galway, L.; Sant'Anna, A.; Verikas, A.; Alonso-Fernandez, F.; Wang, H.; Bond, R. Towards emotion recognition for virtual environments: an evaluation of eeg features on benchmark dataset. *Pers. Ubiquitous Comput.* **2017**, *21*, 1003–1013. doi:10.1007/s00779-017-1072-7.
26. Rosenstein, M.T.; Collins, J.J.; De Luca, C.J. A practical method for calculating largest Lyapunov exponents from small data sets. *Phys. D Nonlinear Phenom.* **1993**, *65*, 117–134. doi:10.1016/0167-2789(93)90009-P.
27. Godin, C.; Prost-Boucle, F.; Campagne, A.; Charbonnier, S.; Bonnet, S.; Vidal, A. Selection of the Most Relevant Physiological Features for Classifying Emotion. Proc. 2nd Int. Conf. Physiol. Comput. Syst., 2015, pp. 17–25. doi:10.5220/0005238600170025.
28. van Gent, P.; Farah, H.; van Nes, N.; van Arem, B. Analysing noisy driver physiology real-time using off-the-shelf sensors: Heart rate analysis software from the taking the fast lane project. *J. Open Res. Softw.* **2019**, *7*, 32. doi:10.5334/jors.241.

29. Dissanayake, V.; Seneviratne, S.; Rana, R.; Wen, E.; Kaluarachchi, T.; Nanayakkara, S. SigRep: Toward Robust Wearable Emotion Recognition with Contrastive Representation Learning. *IEEE Access* **2022**, *10*, 18105–18120. doi:10.1109/ACCESS.2022.3149509.
30. Islam, M.R.; Moni, M.A.; Islam, M.M.; Rashed-Al-Mahfuz, M.; Islam, M.S.; Hasan, M.K.; Hossain, M.S.; Ahmad, M.; Uddin, S.; Azad, A.; Alyami, S.A.; Ahad, M.A.R.; Lio, P. Emotion Recognition from EEG Signal Focusing on Deep Learning and Shallow Learning Techniques. *IEEE Access* **2021**, *9*, 94601–94624. doi:10.1109/ACCESS.2021.3091487.
31. Schuller, B.; Vlasenko, B.; Eyben, F.; Wöllmer, M.; Stuhlsatz, A.; Wendemuth, A.; Rigoll, G. Cross-Corpus acoustic emotion recognition: Variances and strategies. *IEEE Trans. Affect. Comput.* **2010**, *1*, 119–131. doi:10.1109/T-AFFC.2010.8.
32. Zhang, T.; Ali, A.E.; Wang, C.; Hanjalic, A.; Cesar, P. Cornet: Fine-grained emotion recognition for video watching using wearable physiological sensors. *Sensors (Switzerland)* **2021**, *21*, 1–25. doi:10.3390/s21010052.
33. Saganowski, S.; Komoszyńska, J.; Behnke, M.; Perz, B.; Kunc, D.; Klich, B.; Kaczmarek, Ł.D.; Kazienko, P. Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables. *Sci. Data* **2022**, *9*, 158. doi:10.1038/s41597-022-01262-0.
34. Bota, P.; Brito, J.; Fred, A.; Cesar, P.; Silva, H. A real-world dataset of group emotion experiences based on physiological data. *Sci. Data* **2024**, *11*, 116. doi:10.1038/s41597-023-02905-6.
35. Santamaria-Granados, L.; Munoz-Organero, M.; Ramirez-Gonzalez, G.; Abdulhay, E.; Arunkumar, N. Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS). *IEEE Access* **2019**, *7*, 57–67. doi:10.1109/ACCESS.2018.2883213.
36. Cittadini, R.; Tamantini, C.; Scotto di Luzio, F.; Lauretti, C.; Zollo, L.; Cordella, F. Affective state estimation based on Russell's model and physiological measurements. *Sci. Rep.* **2023**, *13*, 9786. doi:10.1038/s41598-023-36915-6.
37. Bustos-López, M.; Cruz-Ramírez, N.; Guerra-Hernández, A.; Sánchez-Morales, L.N.; Cruz-Ramos, N.A.; Alor-Hernández, G. Wearables for Engagement Detection in Learning Environments: A Review. *Biosens. 2022, Vol. 12, Page 509* **2022**, *12*, 509. doi:10.3390/BIOS12070509.