A machine learning model for emotion recognition from physiological signals

J.A. Domínguez-Jiménez, K.C. Campo-Landines, J.C. Martínez-Santos, E.J. Delahoz, S.H. Contreras-Ortiz *

Universidad Tecnológica de Bolívar, Km 1 Vía Turbaco, Cartagena de Indias, Colombia

ARTICLE INFO

ABSTRACT

Emotions are affective states related to physiological responses. This study proposes a model for recognition of three emotions: amusement, sadness, and neutral from physiological signals with the purpose of developing a reliable methodology for emotion recognition using wearable devices. Target emotions were elicited in 37 volunteers using video clips while two biosignals were recorded: photoplethysmography, which provides information about heart rate, and galvanic skin response. These signals were analyzed in frequency and time domains to obtain a set of features. Several feature selection techniques and classifiers were evaluated. The best model was obtained with random forest recursive feature elimination, for feature selection, and a support vector machine for classification. The results show that it is possible to detect amusement, sadness, and neutral emotions using only galvanic skin response features. The system was able to recognize the three target emotions with accuracy up to 100% when evaluated on the test data set.

1. Introduction

Emotions are affective states that influence behavior and cognitive processes. They appear as a result of external or internal stimuli, and are accompanied by physical and physiological reactions. There are several different emotions that can be distinguished from each other by facial expressions, and behavioral and physiological responses [1]. Plutchik proposed a psychoevolutionary theory of emotions that considers eight primary emotional states: fear, anger, joy, sadness, acceptance, disgust, expectancy, and surprise [2]. Other emotions that have been considered in the literature include interest, contempt, guilt, and shame. Emotions can be characterized by two features: valence or pleasantness, and arousal or activation [3]. A two dimensional model of emotions was proposed by Russell [4]. This model organizes emotions according to their valence and arousal, as seen in Fig. 1.

The onset and intensity of emotions have been associated to neural and physiological activity, thoughts and culture. From the physiological point of view, it is considered that emotions are generated by physiological reactions to events [5].

Automatic emotion recognition has been a topic of interest since the last century. Previous works have developed methods for emotion detection from speech features [6,7], facial expressions [8], body gestures [9], and even touches on sensitive screens [10]. As people can hide their emotions or pretend them, methods based only on physical cues may fail to identify the true emotional state of a person. Some studies have used multimodal approaches that combine speech, facial and physiological signals for emotion recognition [11,12].

Changes in physiological signals related to emotional states are involuntary, and people are often unaware of them. Therefore, physiological signal analysis can be a reliable method for emotion recognition. Previous studies have shown that biosensors can be useful in the task of emotion detection by monitoring autonomic nervous system (ANS) activity [13–15].

Another advantage of using biosignals for emotion recognition is that the system can be designed to be wearable and unobtrusive. Since the 1990s, several studies have proposed the use of wearable technologies for emotion detection. In 1997, Picard and Healey introduced the concept of "affective wearables" that are electronic devices equipped with sensors to monitor signals such as galvanic skin response (GSR), blood volume pressure (BVP), heart rate (HR), and electromyogram (EMG), with the purpose of recognizing the wearer's affective states [16]. Later, Scheirer et al. developed a wearable device for facial expression recognition using glasses to sense facial muscle movements to identify expressions such as confusion or interest [17]. In 2004, Haag et al. proposed a system that acquires respiration rate (RSP), electrocardiogram (ECG), GSR,

* Corresponding author.
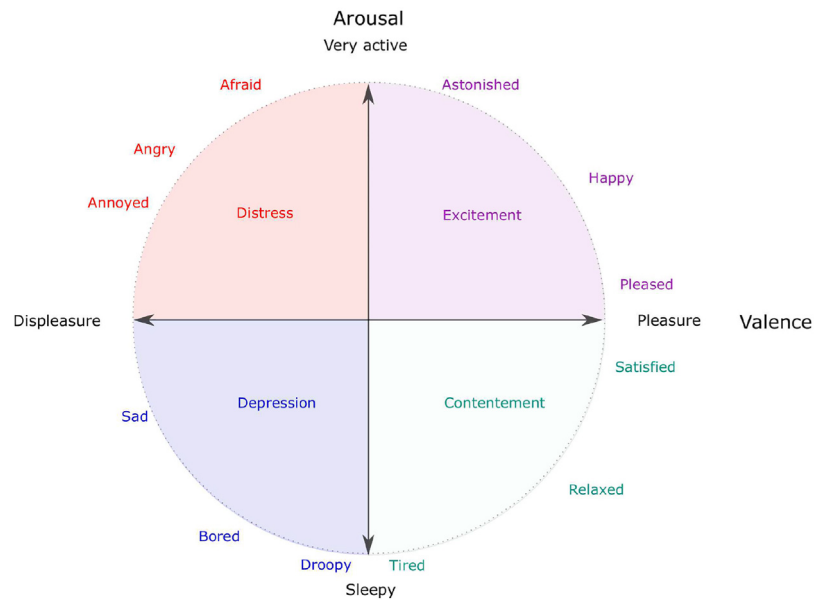 *E-mail address:* scontreras@utb.edu.co (S.H. Contreras-Ortiz).

Fig. 1. Russell's circumplex model of affect [4].

and facial EMG signals and used pattern recognition techniques to recognize emotion valence and arousal, with high accuracy [18]. Although this system uses cabled sensors and it is not wearable, the authors stated that in the future, sensors will become small enough to make the design of wearable devices for emotion recognition possible. Recent studies have proposed frameworks for emotion detection using off-the-shelf wearable sensors, smartphones, and mobile platforms [19,20]. These technologies have also been used in other applications such as health care, education, gaming, and sports [21–23].

In emotion studies, emotion elicitation is an essential and sometimes difficult task. There are different types of stimuli that can be used for emotion elicitation in an experimental setup. The following studies have used pictures, video fragments, or music as stimuli. Gouizi et al. used pictures from the international affective picture system (IAPS) to induce emotions, and recorded EMG, respiratory volume (RV), skin temperature (SKT), skin conductance (SKC), BVP, and HR. Then, they used a support vector machine (SVM) to classify six basic emotions (joy, sadness, fear, disgust, neutrality, and amusement) with a recognition rate of 85% [24]. A more recent work used the Geneva affective picture database (GAPED) for emotion elicitation. They developed a system to classify valence and arousal from PPG and GSR signals with accuracy up to 86.7% for a single-user model [25].

Liu et al. used video fragments to elicit four emotions (happiness, grief, anger, and fear), and recorded GSR. Emotions were classified with a SVM with accuracy of 66.67% [26]. Ayata et al. developed a music recommendation system based on emotions from GSR and photoplethysmography (PPG) signals. They used the database for emotion analysis using physiological signals (DEAP) that includes PPG, GSR, and EMG signals from subjects during video watching, and obtained accuracies up to 72.06% and 71.05% for arousal and valence prediction respectively [27]. Finally, Balasubramanian et al. studied emotional responses to music using electroencephalography (EEG) signals. They compared the valence and arousal of perceived emotions with the valence and arousal of induced ones [28].

This paper proposes an approach for emotion recognition using biosignal processing and machine learning techniques. We developed an instrumented glove with two off-the-shelf sensors to acquire PPG and GSR signals. These biosignals were recorded on volunteers while they were watching video clips for emotion elici-

tation. Signal features were carefully selected, and several machine learning techniques were evaluated. The proposed system is able to identify amusement, sadness, and neutral state with high accuracy. Preliminary results of this research project were published in a conference paper [29]. The purpose of the previous work was to determine the relationship between mean values of PPG and GSR signal features and emotional states. This paper proposes a new experimental protocol for emotion elicitation and biosignal recording and a complete approach for emotion recognition. The rest of the paper is organized as follows. Section 2 describes the stimuli selection, experimental protocol, selected features, and performance measurements. Then, Section 3 presents the results. Finally, Section 4 concludes the paper.

2. Methods

Fig. 2 shows a block diagram of the proposed methodology for emotion recognition. An instrumented glove was developed to acquire two biosignals: PPG and GSR during an emotion elicitation experiment. These signals were analyzed in frequency and time domains to extract a set of features, and the most significant ones were selected to train a classifier. The signal processing tasks were done in Matlab (Mathworks Inc.) and the statistical analysis in RStudio (Version 1.1.442).

The system was designed to identify three emotional states: amusement, sadness, and neutral (non-emotional). Amusement is a feeling that appears as a result of experiencing something funny, and can be located in the first quadrant of the circumplex, with positive valence and arousal. On the other hand, sadness is considered an unpleasant feeling, and is located in the third quadrant of the circumplex, with negative valence and arousal [15]. Finally, neutral can be located in the center of the model.

Below, there is a description of the experimental protocol for emotion elicitation, and data acquisition and processing stages.

2.1. Experimental protocol

We invited 42 healthy subjects aged 18–25 years to participate in the experiments. This study was approved by the Ethics Committee of Universidad Tecnológica de Bolívar. A block diagram of the protocol is shown in Fig. 3.
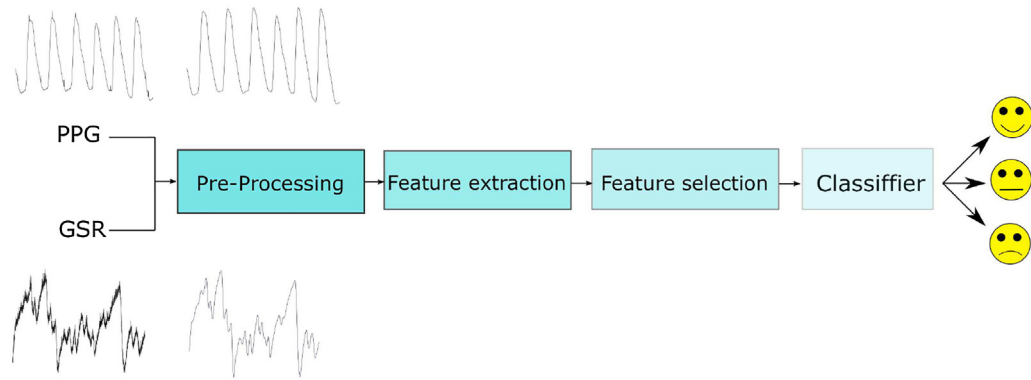
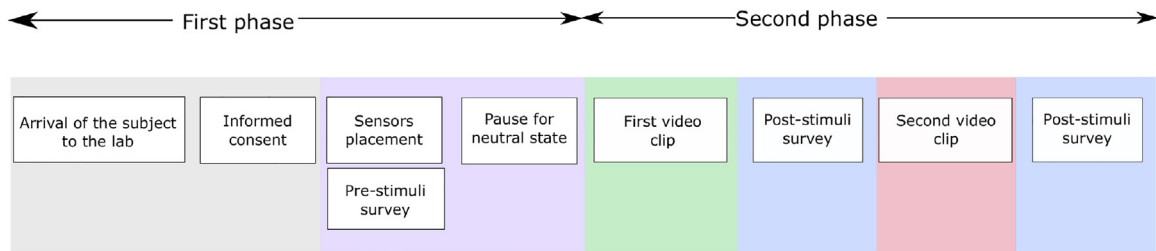**Fig. 2.** Diagram of the data processing stages for emotion classification.

Fig. 3. Experimental protocol. Physiological signals were recorded before and during the presentation of both video clips.
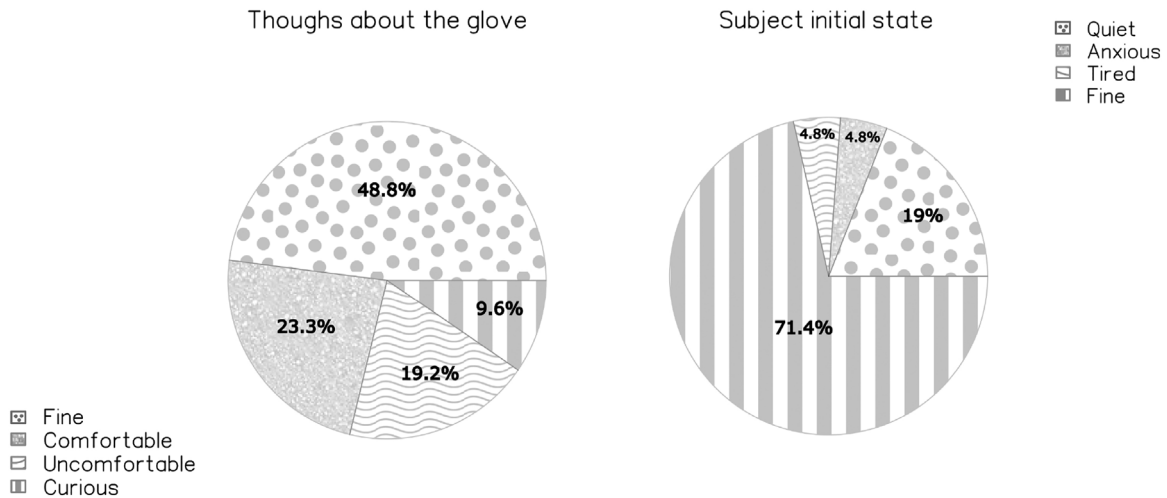

Fig. 4. Results of pre-stimuli survey.

The experiments were conducted in the Bioengineering lab of the university. After the arrival of the subject, a research assistant explained the procedure and answered the subject's questions. He said that the purpose of the study was to analyze physiological signals during the observation of video clips, without mentioning the emotion detection aim, to facilitate spontaneous emotion elicitation. The subject signed an informed consent and the instrumented glove was put on the left hand. Then, the subject was asked to fill in a pre-stimuli survey to know his/her initial emotional state. The results of the survey are shown in Fig. 4. It can be seen that most of the subjects reported feeling fine and comfortable with the glove.

When the subject was observed to be in a neutral state, the signals from the sensors were recorded for two minutes. After that, the subject was asked to watch a video clip that evokes sadness. Biosignals were recorded during the presentation of the stimuli. After the end of the video, the subject was asked to fill in a post-stimuli survey to inquire about his/her emotions. Then, a second video clip that evokes amusement was presented and the procedure was repeated.

Finally, the research assistant took off the glove, answered questions and thanked the subject for his/her participation.

2.2. Stimuli selection

We used video clips from the FilmStim database for emotion elicitation [30]. This database is composed of 70 video clips to evoke several emotions: anger, sadness, fear, disgust, amusement, tenderness and neutral state. We selected two video clips with a duration of approximately 2′40″. Below, there is a brief description of the scenes.

- *Sadness.* The dreamlife of angels. Marie commits suicide by jumping out of a window.
- *Amusement.* When Harry met Sally. Sally fakes an orgasm in a restaurant, provoking Harry's embarrassment.

To confirm the effectiveness of the selected video clips to evoke the target emotions, we performed an analysis of the scores given
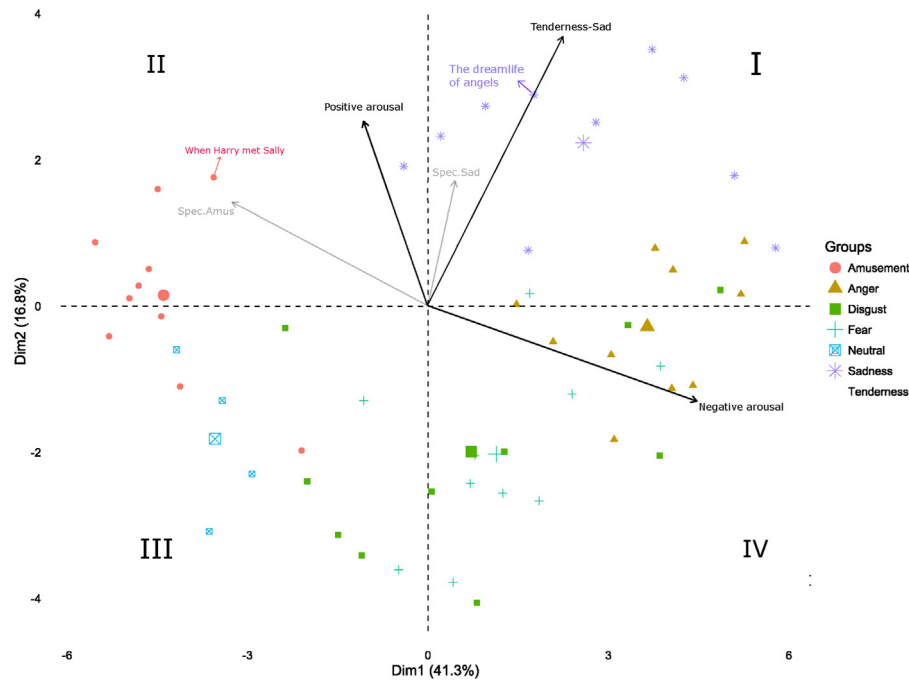
Fig. 5. Exploratory analysis based on principal components of the data provided by Shaefer et al. [30].

by 364 volunteers that rated the clips. The evaluated parameters were subjective arousal, positive and negative effect, emotional discreteness scores, and fifteen mixed feelings scores.

Ten films were selected per emotional category. Due to the high dimensionality (40 dimensions) of the records, we used principal component analysis (PCA) to find a visual representation of the video clips. Fig. 5 shows the two first dimensions which explain 58.01% of the variability of the data. The selected amusement video clip is located in the second quadrant. This location maximizes the discreteness coefficient of amusement and is in the direction of the positive arousal (PA) score. On the other hand, the sadness video clip is located in the first quadrant, in the growing direction of the discreteness coefficient of sadness. This location contributes to the negative arousal (NA) score.

2.3. Physiological signals acquisition

The PPG and GSR signals were selected to assess the subject's emotional states. Below, there is a description of these signals.

- GSR: the galvanic skin response is a measure of the electrical conductivity of skin. It varies with changes in the activity of sweat glands controlled by the ANS [31–33]. Previous studies have shown that skin conductivity increases monotonically with emotional arousal [34,35]. We used a commercial sensor (Grove-GSR sensor) to measure the GSR signal. The sensor was attached to the subject's middle and index fingers.
- PPG: photoplethysmography is a non-invasive optical technique that detects blood volume changes in tissues. It can be used to measure heart rate and oxygen saturation. Heart rate is regulated by the sympathetic and parasympathetic nervous system and varies with emotional states. Psychological arousal is characterized by an increased heart rate [36]. Additionally, it has been observed that heart rate decelerates due to unpleasant stimuli [35]. In this study, we measured heart rate with a commercial PPG sensor (Gravity, DFRobot) fastened to the subject's ring finger.

An instrumented glove was developed to hold the data acquisition system (see Fig. 6) [29]. The sensors were connected to a

Fig. 6. Instrumented glove for signal acquisition.

microcontroller board (Bluno Nano, DFRobot). The sampling frequency was 500 Hz and the data was acquired through the USB port.

2.4. Feature extraction

The recordings from 37 subjects had acceptable quality and were analyzed. The signals from other five subjects were saturated and were excluded from the study.

Previous to feature extraction, the signals were filtered to reduce noise. The PPG signal was processed with a 100th-order band pass FIR filter with corner frequencies equal to 0.1 Hz and 10 Hz. The GSR signal was low pass filtered with a 1000th-order FIR filter with a corner frequency of 1 Hz.

Heart rate was estimated from the PPG signal using short-time fast Fourier transform (ST-FFT). We used 5-s length windows and zero-padding to obtain an effective frequency resolution of two beats per minute (BPM).

Table 1
Feature selection methods and final predictors.

Method	Number of predictors	Predictors
RF-RFE	4	scraonv,scrpnv,scravd,crm4
GA	5	scravd,scraonv,scrpnv,crm3,hrstd
SW-F	8	scravd,scrpnv,scrstd,crm2,scrdr crm1,thd5,LFHF
SW-Bidir	16	scravd,scrstd,scrdr,scraonv,scrpnv,crm1,crm3,crm4 hrm,hrstd,hrdr,hrssdn,hrrmssd,thd2,thd3,thd5

The features extracted from the signals were selected according to previous studies in emotion recognition [37,27,38], and considering related metrics and norms used in healthcare [39,40].

2.4.1. Photoplethysmography signal
The PPG signal was analyzed in time and frequency domain to obtain 13 features.

- Time domain features: root mean square differences of successive R–R intervals (HRRMSSD) and standard deviation of normal to normal R–R intervals (HRSDNN)
- Frequency domain features: mean of heart rate (hrmean), standard deviation of heart rate (hrstd), heart rate dynamic range (hrdr), heart rate mode (hrmode), harmonic distortion of the second, third, fourth and fifth harmonics of the PPG signal (THD2, THD3, THD4, THD5). The power spectrum of the PPG signal was divided into two bands: low frequency (0.04–0.15 Hz) and high frequency (0.15–0.5 Hz) to compute the normalized power in the low and high frequency bands (LFnu and HFnu respectively), and the ratio of LF to HF power (LFHFnu).

2.4.2. Galvanic skin response signal
The GSR signal was analyzed in time domain, to characterize its variability. A total of 14 features were calculated: mean (scrmean), standard deviation (scrstd), dynamic range (scrdr), mean of the derivative (scravd), negative values of the derivative (scraonv), and ratio of negative values over the total number of samples (scrpnv). To obtain information about nonlinear and non-stationary components, empirical mode decomposition (EMD) was used. We limited the number of modes to four, and calculated the energy (emf1, emf2, emf3, emf4) and the zero crossing rate (crm1, crm2, crm3, crm4) of each mode.

2.5. Feature selection

We computed a total of 27 features, but not all of them were considered for the classification stage, because possible dependencies among the variables may decrease the classifier performance. There are several feature selection methods that have been used in the literature, and some of them are mentioned next. Chih-Fong used *T*-test, correlation matrix, stepwise regression (SW), principle component analysis (PCA) and factor analysis (FA) to select the most representative features for bankruptcy prediction [41]. Niu et al. used genetic algorithms (GAs) and K-neighbors for feature selection for emotion recognition from physiological signals [42]. Zvarevashe et al. used the random forest recursive feature elimination (RF-RFE) algorithm with gradient boost machines (GBMs) for gender voice recognition [43].

To find the optimal feature subset to represent the target emotions, we used several feature selection techniques including SW, RF-RFE, and GAs. Final predictors for each technique are presented in Table 1.

2.5.1. Stepwise regression
This method consists of regressing multiple variables by removing the least contributing predictors step by step. Only independent variables with non-zero coefficients are included into the final regression model. There are three types of SW: forward selection (FW), backward selection (BW), and bidirectional elimination (BIDIR). We used SW with the Akaike information criterion (AIC) as stop criterion.

2.5.2. Genetic algorithms
GAs are inspired in natural evolution. In nature, organisms have evolved over generations to better adapt to their environment. GAs can be used to maximize the performance of a predictive model on an unseen data set avoiding overfitting. GAs need a population of individuals and generations to produce better approximations depending on some parameters such as mutation and crossover probability. At each generation, according to a fitness criterion, a new set of individuals, i.e. subsets of predictors, is created and also recombined using operators from natural genetics. Among the fitness measures are root mean squared error (RMSE) and classification accuracy [44,45]. In this work, 10-fold cross-validation was implemented as resampling method, with classification accuracy as fitness measure. Only internal performance, provided by the training data, was used in the search. External performance was not used. The best performance was achieved at iteration 26 with a subset of 5 variables.

2.5.3. Random forest – recursive feature elimination
Resampling methods such as cross-validation and bootstrap are useful for feature selection during model building. These methods can maximize the model's performance but the computational cost increases. RF-RFE provides a reliable assessment of predictors and presents a ranked set of the best predictors at the end. In this work, 10-fold cross-validation was used as resampling method, and RF-RFE was used to predict during each resample. The best performance was empirically obtained with four predictors.

2.6. Classification

The selected features were used as input to the classification stage. Several models were evaluated for the purpose of recognition of three emotional states: amusement, sadness and neutral.

2.6.1. Data preparation
The data set was standardized and centered. Then, it was divided into 80% for model training, and 20% for testing the model with unseen data. The data were balanced during training and test, i.e. we used the same number of observations of each class.

2.6.2. Classification models
We used several classification methods that include support vector machines (SVM), linear discriminant analysis (LDA and SLDA), multinomial regression (MN), decision trees (DT), and naive Bayes (NB). We also evaluated ensemble models, such as extreme gradient boost (XGBTREE), boosted logistic regression (BLR), lasso and elastic-net regularized generalized linear models (GLMNET) and bagging trees (TBAG). These classifiers were trained and tested over the unseen data with a 10-folds cross-validation algorithm to obtain subject-dependent models for emotion recognition.

2.7. Performance metrics

The performance metrics used in this work are accuracy and the receiver operating characteristic (ROC) curve. The ROC curve considers the number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), and area under the curve

Table 2
One-way ANOVA results for statistically significant predictors.

Predictor	df	Sum sq	Mean sq	F	p
crm4	2	7.41e−07	3.70e−07	14.268	4.37e−06**
crm3	2	1.66e−06	8.30e−07	11.456	3.83e−05**
crm2	2	4.92e−06	2.46e−06	9.8188	1.42e−04**
crm1	2	2.68e−05	1.34e−05	7.8087	7.60e−04**
scravd	2	5.78e−10	2.89e−10	24.572	3.46e−09**
scraonv	2	980,257,470	490,128,735	10.563	7.821e−05**
scrpnv	2	0.42915	0.214576	22.530	1.305e−08**
emf4	2	3.42e−06	1.71e−06	3.4093	3.75e−02*
hrstd	2	48.51	24.25	2.78	0.06.

Signif. codes: 0 ** 0.01 * 0.05.

(AUC) to illustrate the detection ability of a model. These metrics were computed during a cross-validation procedure for each model. Below, there is a description of the metrics.

2.7.1. Accuracy

It computes the amount of TP and TN over the total of observations:

$$Accuracy = \frac{TP + TN}{P + N}$$

2.7.2. Receiver operating characteristic

This metric describes the TP percentage versus the FP percentage. It helps to understand how sensitive (TP rate) and specific (TN rate) is a model. The ROC curve can be obtained by plotting the TP rate against FP. The best possible AUC is 1.0. The diagonal line in the ROC depicts randomness.

3. Results and discussion

In this section, we present and discuss the results of the statistical analysis, classification, and post-stimuli surveys.

3.1. Statistical analysis

The hypothesis of this analysis is to prove that at least one pair of emotions varies with respect to the mean levels in at least one predictor. Table 2 shows the one-way ANOVA results for the predictors that are statistically significant with 95% certainly. The analysis of GSR features including scravd, scrpnv and scraonv shows that by computing just derivative characteristics, a statistically significant difference can be reached for emotion recognition. EMD features are also relevant, since the zero crossing rates and energy values of the four modes are statistically significant. On the other hand, hrstd is the only PPG feature that provides differences on each target emotion. However, it does not provide a statistically significant difference since its p-value is above the limit. Hence, it is possible to detect emotional changes only from GSR predictors even in their mean levels with 95% certainly. The reason why HR features were not significant to identify among amusement, sadness and neutral may be that these emotions have similar values of arousal, and HR is a signal that mostly varies with changes in this attribute.

3.2. Cross-validation

This section presents the 10-fold cross-validation[1] results listed in Table 3 for each subset of predictors presented in Table 1. Table 4 shows the tuning hyperparameters for all the models and the selected ones.

The worst subset of predictors is the SW-FW subset. The best mean accuracy for this data set was reached by GLMNET with 64%,

---

[1] Seed was fixed at 10 for reproducibility of results.

Table 3
10-Fold cross-validation results for each model.

Model	RFE						GA						SW BIDIR						SW FW					
	Accuracy			ROC			Accuracy			ROC			Accuracy			ROC			Accuracy			ROC		
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
SVML	0.88	0.97	1.00	0.90	0.99	1.00	0.88	0.95	1.00	0.92	0.99	1.00	0.77	0.90	1.00	0.87	0.97	1.00	0.44	0.57	0.66	0.62	0.75	0.88
SVMR	0.77	0.85	1.00	0.87	0.96	1.00	0.66	0.84	1.00	0.85	0.83	1.00	0.55	0.73	1.00	0.79	0.89	1.00	0.44	0.63	0.77	0.66	0.78	0.92
KNN	0.66	0.81	1.00	0.78	0.89	1.00	0.44	0.75	0.88	0.73	0.86	0.93	0.44	0.66	0.88	0.59	0.78	0.91	0.44	0.6	0.88	0.58	0.74	0.92
LDA	0.77	0.86	1.00	0.88	0.96	1.00	0.77	0.83	1.00	0.88	0.96	1.00	0.77	0.87	1.00	0.77	0.93	1.00	0.44	0.54	0.66	0.53	0.73	0.87
SLDA	0.66	0.83	1.00	0.85	0.94	1.00	0.66	0.84	1.00	0.85	0.94	1.00	0.66	0.84	1.00	0.85	0.94	1.00	0.22	0.57	0.77	0.53	0.76	0.94
MN	0.88	0.98	1.00	1.00	1.00	1.00	0.88	0.97	1.00	0.98	0.99	1.00	0.77	0.82	1.00	0.85	0.95	1.00	0.44	0.57	0.77	0.62	0.79	0.94
NB	0.44	0.7	0.88	0.61	0.82	0.98	0.44	0.65	0.88	0.64	0.81	0.94	0.44	0.58	0.66	0.53	0.73	0.83	0.33	0.56	0.66	0.46	0.74	0.90
DT	0.22	0.62	0.77	0.51	0.76	0.86	0.22	0.60	0.77	0.48	0.72	0.90	0.22	0.58	0.77	0.51	0.73	0.90	0.44	0.52	0.66	0.55	0.66	0.77
XGBTREE	0.66	0.80	1.00	0.74	0.94	1.00	0.44	0.8	1.00	0.75	0.89	1.00	0.55	0.74	0.88	0.75	0.87	0.94	0.44	0.64	0.88	0.72	0.80	0.94
GLMNET	0.88	0.98	1.00	0.90	0.99	1.00	0.88	0.98	1.00	0.92	0.99	1.00	0.77	0.96	1.00	0.92	0.99	1.00	0.55	0.64	0.77	0.62	0.81	0.92
BOOSTLR	0.57	0.82	0.88	0.79	0.92	0.98	0.62	0.85	1.00	0.77	0.90	1.00	0.5	0.73	0.85	0.77	0.87	1.00	0.22	0.58	1	0.44	0.75	0.90
TBAG	0.55	0.76	0.88	0.76	0.9	1.00	0.55	0.73	0.88	0.77	0.88	0.94	0.55	0.66	0.77	0.63	0.83	0.92	0.33	0.56	0.77	0.62	0.78	0.89

Table 4
Classification models tuning and final parameters.

Classification algorithm	Tuning parameters	Final model
KNN	k-neighbors = c(1:50)	17 K-Neighbors
MN	decay = c(0,1e−04,1e−01)	decay = 1e-04
GLMNET	α = c(0,1), λ = seq(0.001,0.1 by 0.001)	α = 1, λ = 0.06
SVML	cost = 2^c(0:5)	c = 8, Support Vectors = 34
SVMR	σ = 2^c(-25,-20,-15,-10,-5,0), cost = 2^c(0:5)	c = 32, Support Vectors = 47, σ = 0.03125
DT	cp = 2^c(-32,-25,-20,-15,-10,-5,-2,0)	cp = 0.02643
BLR	nIter = 50	nIter = 11
NB	fL = c(0,0.5,1.00), bw_adjust = c(0,0.5,1.0)	fL = 1, bw_adjust = 1.0
XGBTREE	n.trees = 500, max_depth = c(1:4), η = c(0.01,0.1)	n.trees = 500, max_depth = 3, η = 0.1
LDA	No parameters need for this model	
SLDA	Both directions	
TBAG	No parameters need for this model	

with a minimum value of 55% and a maximum of 77%.[2] SVM with a radial kernel (SVMR) reached the same mean accuracy but its variability is higher. With respect to the ROC curve, the maximum ROC value corresponds to GLMNET. The BLR model was the only one that did not reach the expected performance with respect to its variance in ROC, despite its mean accuracy was 58.21%.

The results for the SW-BIDIR subset show that the model that maximizes the mean accuracy and the ROC is GLMNET, with 96% and 99% respectively. Similarly, SVM with linear kernel (SVML) reached a mean accuracy of 90% within the same variability of the aforementioned. In comparison with the SW-F subset, this one improves the global mean accuracy, from 58% to 76.92% and the global mean ROC from 76.28% to 87.93%. The results for this subset also show that the MN, LDA, and SLDA models are able to discriminate among the target emotions with high mean accuracy. The minimum ROC in the worst case scenario corresponds to LDA with 77%. The models KNN, NB, and DT have the lowest accuracies. Finally, all the models are able to generate information over randomness, since the lowest ROC is 51%.

The GAs subset performance is better than the SW-BIDIR. The mean global accuracy was 82.04% and the ROC 90.95%. Particularly, there is no a statistically significant difference among the mean accuracy levels of GLMNET, MN, and SVML models. GLMNET is the model that maximizes the mean accuracy to 98% and MN is the one that maximizes the ROC value, since its variability goes from 98% to 100%. In addition, once again the NB, KNN and DT models are not able to predict the target emotions within a 95% of certainly.

The RFE subset is the one that provides the best global performance with mean global accuracy of 83.61% and mean global ROC of 92.36%. Results are shown in Fig. 7. The results in Table 3 show that the models with the best performance are GLMENT, MN and SVML. Fig. 8 shows the ROC curve for each model. This figure is helpful to identify the best model to discriminate among emotion pairs i.e. neutral-sadness (N-S), amusement-sadness (A-S), and neutral-sadness (N-S). There are three ROC plots, one per emotion. We can state that all models are able to identify amusement when it is the target. GLMNET could fail in identify amusement when the pair is A-S. On the other hand, when sadness is the target, MN and SVML may fail if the emotion pair is A-N. Once again, GLMNET has problems with the A-S pair. Finally, when neutral is the target, all models have the same behavior regarding to A-N and N-S. They are able to discriminate neutral when it is present. In contrast, when the pair is A-S, the models tend to get confused as they lie close to the diagonal in the ROC curve. Table 5 presents the mean AUC values for each case. Since the priority of this work is to discriminate between amusement and sadness, over neutral, SVML is the

---

Table 5
AUC values per emotion target and model.

Model	Target emotion	AUC
MN	A	0.93
	N	0.89
	S	0.88
SVML	A	0.96
	N	0.86
	S	0.91
GLMNET	A	0.92
	N	0.89
	S	0.86

best model to classify the emotions when the target is amusement (96%) or sadness (91%).

Finally, the information of the SVML model's performance during training and test is presented by the confusion matrices in Fig. 9, where the number in parentheses represents the classified observations per class. When evaluated on the test data, the SVML was able to recognize emotional states with mean accuracy of 100%. When evaluated on the training data, the model identified the target emotions with mean accuracy of 97.78%. This model has a no-information rate of 33.33% with the unseen data.

3.3. Post-stimuli survey

The self-report survey had five questions that go from general to particular to track the emotion elicitation process. These questions are listed below:

1. What impressions do you have about what you have just seen and experienced?
2. Did you experience any change in your emotional state?
3. Did you experience something positive, negative or something else?
4. Express in a single word what you felt.
5. On a scale from 1 to 5 (where 1 is little amused/sad and 5 very amused/sad) how would you rate the intensity of what you felt?

The answers to the survey show that most of the subjects that watched the amusement video clip declared experiencing the target emotion. All subjects stated feeling a change in their emotional state. From the valence domain, 86.50% declared feeling a positive emotion, and nobody said that felt a negative emotion. All subjects described their emotional state as fun and amusement, and their facial expressions were according to it. Fig. 10 and Table 6 show these results.

In the case of the sadness video clip, 56.75% of the population stated feeling a change in their emotional state. From the valence domain, 37.83% of the subjects declared feeling a negative emotion, 56.75% something else, and the rest a positive emotion. Regarding

---

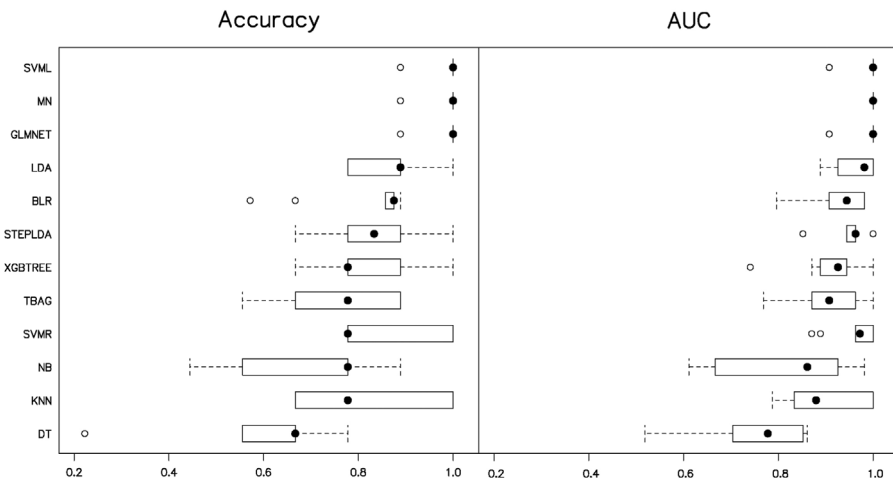[2] These intervals denote the variance of the 10 folds cross-validation results with 95% certainly.

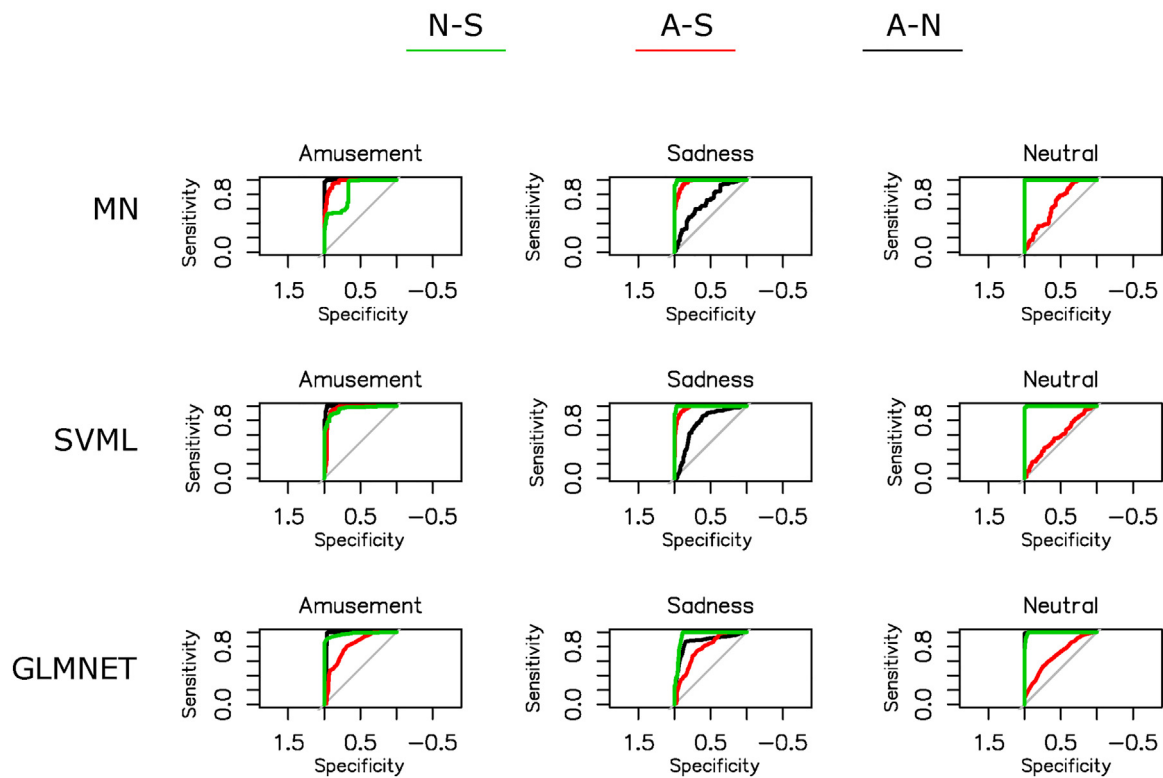**Fig. 7.** Results for both metrics on the final subset selected.

Fig. 8. ROC curve for the best three models with the selected subset.

Table 6
Analysis of the survey responses for the amusement video.

Reported valence	Reported emotion	Intensity of amusement	Subjects
Positive	Amusement, laugh, nothing, satisfaction	5	19
	Amusement, laugh	4	9
	Amusement	3	4
	Amusement, happiness	5	3
Nor positive	Amusement	4	1
Nor negative	Boring	3	1

the main evoked emotion, as seen in Fig. 10, only two participants (5.4%) reported sadness, and other three (8.1%) reported similar emotions: shame and loneliness. Most of the subjects reported feeling surprise. In the answers to the first question, ten subjects used words related to sadness such as sad, sadness, sorrow, tragic,

compassion, grief, depression, empathy, and touched. Finally, with respect to the last question, 40.5% of the subjects rated the intensity of sadness as 2, 18.9% as 3, and 5.4% as 4 (see Table 7). Therefore, 64.9% of the subjects stated feeling moderate to intense sadness.
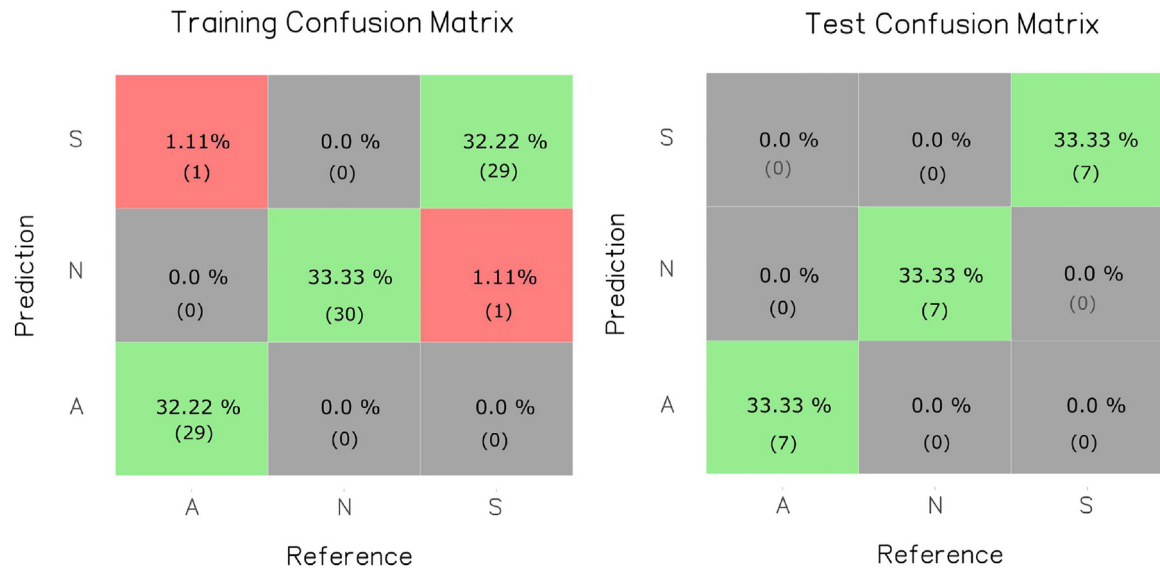
Fig. 9. Confusion matrices for the training (left) and test (right) samples for the SVML model.

Table 7
Analysis of the survey responses for the sadness video.

| Reported valence	Reported emotion	Intensity of sadness	Subjects
Negative	Sadness, fear, astonishment, impression	3–4	6
	Sorrow, surprise, fear	2	6
	Intrigue, fear	1	2
Nor positive	Surprise, impression	3	3
Nor negative	Astonishment, bewilderment, loneliness, surprise, intrigue, suspense	2	8
	Astonishment, bewilderment, apathy, surprise, anxiety, impression, curiosity, nothing	1	10
Positive	Tension	2	1
	Suspense	1	1
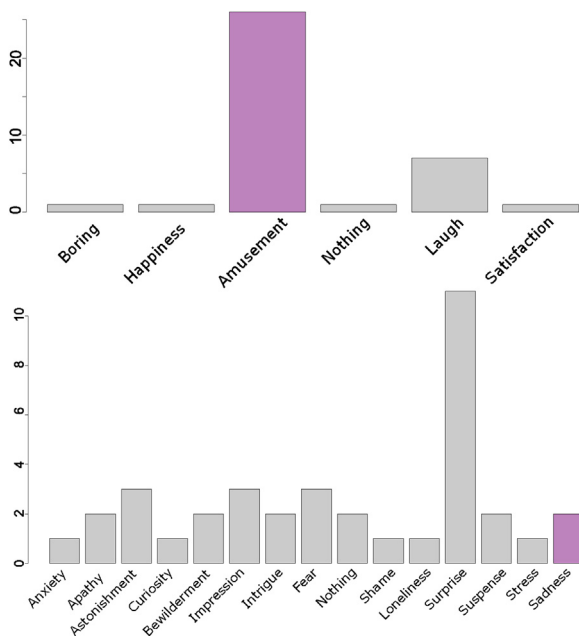
Fig. 10. Main emotion evoked by the amusement video (top) and by the sadness video (bottom) as reported by the subjects in the post-stimuli survey.

It can be seen that the elicitation of the sadness emotion was more difficult compared to the amusement emotion, as well as the analysis of the survey responses to identify the evoked emotions. The video clip for sadness elicitation shows a woman that kills herself unexpectedly. This occurs 22 s before the end of the video and right after watching the video, the participants answered the survey. It can be expected that due to its intensity and closeness to the application of the survey, the subjects would express the surprise as the prevailing emotion in that moment, over other emotions felt before and after the surprise emotion. Surprise is known as one of the briefest emotions (it lasts just a few seconds) and usually mixes with other basic emotions such as joy and sadness once the individual understands what is happening [46].

According to Reeve, people often tend not to express certain emotions that make them feel physically uncomfortable [47]. Furthermore, there are rules for the manifestation of emotions such as neutralization and masking which are learned socially and their regulation is given according to social situations. This could explain the variability and some apparent inconsistencies in the subjects' responses. Considering that the experiments were developed in an artificial context under the presence of the examiner, and that sadness is often considered as an expression of vulnerability or weakness, the subjects may have found difficult to recognize and manifest it openly [48]. However, the responses to the first question shows that sadness could appear through empathy and compassion, and most of the subjects rated the intensity of sadness as moderate to intense in the last question.

3.4. Methodology validation

To validate the proposed methodology, we used the DEAP data set.[3] Since the procedure for video clip selection in our approach was label-based, only videos with labels similar to "amusement"

---

[3] https://www.eecs.qmul.ac.uk/mmv/datasets/deap/.

(joy, happy, fun) and "sadness" (depressing, sadness) were considered. Four videos were selected for "happy" and two for "sadness" per subject. Then, the same features extracted in our approach (except THD1, THD2, THD3, and THD4) were computed for each video clip in all subjects (32). Heart rate features were computed only in time domain because the low sampling frequency (128 Hz) of the signals leads to low resolution in frequency-domain. Due to problems with the convergence of the EMD algorithm, some observations were discarded. Finally, a binary imbalanced subset was obtained with 84 observations for "happy" and 44 observations for "sadness". Data was divided in 80% for training and 20% for testing to obtain a subject-dependent model using the proposed methodology. In this case, we chose metrics as the F1 score that are more informative when data sets are imbalanced. F1 score assesses the model's performance based on the ability to recognize TPs and TNs from the cost associated to recognize each of them [49,50].

As with our data set, the RFE subset (hrmode, scrdr, hrdr, hrssdn, HFnu) is the one with the best global performance. The models that showed the best global performance were the tree-based models. The model that maximized the F1 score was TBAG with 81%. This model also maximized emotion recognition with a mean accuracy of 74.5% (60.05%, 90.32%) with a confidence interval of 95%. The work by Ayata et al. used the same data set for arousal and valence prediction from GSR and PPG signals, and obtained accuracy rates of 72.06% and 71.05% respectively [27]. Our results show that even with the limitations described above, the proposed methodology can be used to design a system for emotion recognition using a different data set.

4. Conclusions

This work shows that emotion recognition is possible from the PPG and GSR signals with high accuracy. The feature selection techniques that were able to maximize the model performance with a less number of predictors were GAs and RFE. We found that derivative features of GSR, and energies and zero crossing rates of its EMD modes allow to correctly classify the target emotional states. For amusement and sadness recognition, the features evaluated on the PPG signal showed not to be significant. This was observed in the ANOVA results too.

Several classification models were trained to select the one that maximizes the accuracy and ROC. Most of the models showed good performance over all the subsets except SW-FW. SVML was the method that provided the best classification performance regarding to its mean accuracy and ROC, in particular to identify amusement and sadness.

Even though much progress has been made in methods to detect emotional states, it is still necessary to work on the identification of effective stimuli to elicit emotions that are strongly shaped by cognitive aspects such as expectations and perceptions, by processes of socialization, personal and cultural history, as in the case of sadness and anger. It is often found that, although the basic emotions are universal, the stimuli that elicit them with high intensity are not. Therefore, it is suggested to implement protocols in which each subject actively and intentionally participates in the selection of what generates emotion, is informed of the intention of the study and helps in the process of evoking his/her own emotion.

Future work includes the evaluation of other emotion elicitation protocols with a higher number of subjects, the addition of other emotions in the analysis, and the development of a wearable system for emotion detection in real time.

Acknowledgments

Declaration of Competing Interest

None declared.

References

[1] P. Ekman, An argument for basic emotions, Cogn. Emot. 6 (3–4) (1992) 169–200.
[2] K.R. Scherer, P. Ekman, Approaches to Emotion, Psychology Press, 1984.
[3] P.J. Lang, The emotion probe: studies of motivation and attention, Am. Psychol. 50 (5) (1995) 372.
[4] J.A. Russell, A circumplex model of affect, J. Pers. Soc. Psychol. 39 (6) (1980) 1161.
[5] W. James, What is an emotion? Mind 9 (34) (1884) 188–205.
[6] R. Tato, R. Santos, R. Kompe, J.M. Pardo, Emotional space improves emotion recognition, Seventh International Conference on Spoken Language Processing (2002).
[7] R. Cowie, R.R. Cornelius, Describing the emotional states that are expressed in speech, Speech Commun. 40 (1–2) (2003) 5–32.
[8] M.A. Turk, A.P. Pentland, Face recognition using eigenfaces, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91, IEEE, 1991, pp. 586–591.
[9] A.P. Atkinson, M.L. Tunstall, W.H. Dittrich, Evidence for distinct contributions of form and motion information to the recognition of emotions from body gestures, Cognition 104 (1) (2007) 59–72.
[10] A. Heraz, M. Clynes, Recognition of emotions conveyed by touch through force-sensitive screens: observational study of humans and machine learning techniques, JMIR Mental Health 5 (3) (2018), e10104, http://dx.doi.org/10.2196/10104.
[11] N. Sebe, I. Cohen, T.S. Huang, Multimodal emotion recognition, in: Handbook of Pattern Recognition and Computer Vision, World Scientific, 2005, pp. 387–409.
[12] M. Soleymani, M. Pantic, T. Pun, Multimodal emotion recognition in response to videos, IEEE Trans. Affect. Comput. 3 (2) (2012) 211–223.
[13] O. Alaoui-Ismaïli, O. Robin, H. Rada, A. Dittmar, E. Vernet-Maury, Basic emotions evoked by odorants: comparison between autonomic responses and self-evaluation, Physiol. Behav. 62 (4) (1997) 713–720.
[14] R.W. Levenson, L.L. Carstensen, W.V. Friesen, P. Ekman, Emotion, physiology, and expression in old age, Psychol. Aging 6 (1) (1991) 28.
[15] I.C. Christie, B.H. Friedman, Autonomic specificity of discrete emotion and dimensions of affective space: a multivariate approach, Int. J. Psychophysiol. 51 (2) (2004) 143–153.
[16] R.W. Picard, J. Healey, Affective wearables, Pers. Technol. 1 (4) (1997) 231–240.
[17] J. Scheirer, R. Fernandez, R.W. Picard, Expression glasses: a wearable device for facial expression recognition, in: CHI'99 Extended Abstracts on Human Factors in Computing Systems, ACM, 1999, pp. 262–263.
[18] A. Haag, S. Goronzy, P. Schaich, J. Williams, Emotion recognition using bio-sensors: first steps towards an automatic system, in: Tutorial and Research Workshop on Affective Dialogue Systems, Springer, 2004, pp. 36–48.
[19] T. Hui, R. Sherratt, Coverage of emotion recognition for common wearable biosensors, Biosensors 8 (2) (2018) 30, http://dx.doi.org/10.3390/bios8020030.
[20] C. Bailon, M. Damas, H. Pomares, D. Sanabria, P. Perakakis, C. Goicoechea, O. Banos, Intelligent monitoring of affective factors underlying sport performance by means of wearable and mobile technology, Proceedings 2 (19) (2018) 1202, http://dx.doi.org/10.3390/proceedings2191202.
[21] K.A. Popat, P. Sharma, Wearable computer applications a future perspective, Int. J. Eng. Innov. Technol. 3 (1) (2013) 213–217.
[22] S. Jhajharia, S. Pal, S. Verma, Wearable computing and its application, Int. J. Comput. Sci. Inform. Technol. 5 (4) (2014) 5700–5704.
[23] I.C. Jeong, D. Bychkov, P. Searson, Wearable devices for precision medicine and health state monitoring, IEEE Trans. Biomed. Eng. (2018) 1, http://dx.doi.org/10.1109/TBME.2018.2871638.
[24] K. Gouizi, F. Bereksi Reguig, C. Maaoui, Emotion recognition from physiological signals, J. Med. Eng. Technol. 35 (6–7) (2011) 300–307, http://dx.doi.org/10.3109/03091902.2011.601784.
[25] G. Udovičić, J. erek, M. Russo, M. Sikora, Wearable emotion recognition system based on GSR and PPG signals, in: Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care, ACM, 2017, pp. 53–59.
[26] M. Liu, D. Fan, X. Zhang, X. Gong, Human emotion recognition based on galvanic skin response signal feature selection and SVM, in: 2016 International Conference on Smart City and Systems Engineering (ICSCSE), IEEE, Hunan, China, 2016, pp. 157–160, http://dx.doi.org/10.1109/ICSCSE.2016.0051.
[27] D. Ayata, Y. Yaslan, M.E. Kamasak, Emotion based music recommendation system using wearable physiological sensors, IEEE Trans. Consumer Electron. 64 (2) (2018) 196–203, http://dx.doi.org/10.1109/TCE.2018.2844736.
[28] G. Balasubramanian, A. Kanagasabai, J. Mohan, N.G. Seshadri, Music induced emotion using wavelet packet decomposition – An EEG study, Biomed. Signal Process. Control 42 (2018) 115–128, http://dx.doi.org/10.1016/j.bspc.2018.01.015.

[29] J. Domínguez-Jiménez, K. Campo-Landines, J. Martínez-Santos, S. Contreras-Ortiz, Emotion detection through biomedical signals: a pilot study, in: 14th International Symposium on Medical Information Processing and Analysis, vol. 10975, International Society for Optics and Photonics, 2018, p. 1097506.

[30] A. Schaefer, F. Nils, X. Sanchez, P. Philippot, Assessing the effectiveness of a large database of emotion-eliciting films: a new tool for emotion researchers, Cogn. Emot. 24 (7) (2010) 1153–1172.

[31] C. Collet, E. Vernet-Maury, G. Delhomme, A. Dittmar, Autonomic nervous system response patterns specificity to basic emotions, J. Auton. Nerv. Syst. 62 (1–2) (1997) 45–57, http://dx.doi.org/10.1016/S0165-1838(96)00108-7.

[32] R.A. McCleary, The nature of the galvanic skin response, Psychol. Bull. 47 (2) (1950) 97.

[33] A. Mundy-Castle, B. McKiever, The psychophysiological significance of the galvanic skin response, J. Exp. Psychol. 46 (1) (1953) 15.

[34] P.J. Lang, M.M. Bradley, B.N. Cuthbert, A motivational analysis of emotion: reflex–cortex connections, Psychol. Sci. 3 (1) (1992) 44–49.

[35] M.M. Bradley, P.J. Lang, Affective reactions to acoustic stimuli, Psychophysiology 37 (2) (2000) 204–215.

[36] B.M. Appelhans, L.J. Luecken, Heart rate variability as an index of regulated emotional responding, Rev. Gen. Psychol. 10 (3) (2006) 229.

[37] D.D. Ayata, Y. Yaslan, M. Kamaşak, Emotion recognition via galvanic skin response: Comparison of machine learning algorithms and feature extraction methods, Istanbul Univ.-J. Electr. Electron. Eng. 17 (1) (2017) 3147–3156.

[38] H.-K. Chen, Y.-F. Hu, S.-F. Lin, Methodological considerations in calculating heart rate variability based on wearable device heart rate samples, Comput. Biol. Med. 102 (2018) 396–401.

[39] F. Shaffer, J. Ginsberg, An overview of heart rate variability metrics and norms, Front. Public Health 5 (2017) 258.

[40] W. von Rosenberg, T. Chanwimalueang, T. Adjei, U. Jaffer, V. Goverdovsky, D.P. Mandic, Resolving ambiguities in the lf/hf ratio: lf-hf scatter plots for the categorization of mental and physical stress from hrv, Front. Physiol. 8 (2017) 360.

[41] C.-F. Tsai, Feature selection in bankruptcy prediction, Knowl.-Based Syst. 22 (2) (2009) 120–127.

[42] X. Niu, L. Chen, Q. Chen, Research on genetic algorithm based on emotion recognition using physiological signals, in: 2011 International Conference on Computational Problem-Solving (ICCP), IEEE, 2011, pp. 614–618.

[43] K. Zvarevashe, O.O. Olugbara, Gender voice recognition using random forest recursive feature elimination with gradient boosting machines, in: 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), IEEE, 2018, pp. 1–6.

[44] M. Gen, R. Cheng, Genetic Algorithms and Engineering Optimization, vol. 7, John Wiley & Sons, 2000.

[45] D.E. Goldberg, J.H. Holland, Genetic algorithms and machine learning, Mach. Learn. 3 (2) (1988) 95–99.

[46] F.K. Miguel, Psicologia das emoç oes: uma proposta integrativa para compreender a express ao emocional, Psico-usf 20 (1) (2015) 153–162.

[47] J. Reeve, Understanding Motivation and Emotion, 1993, pp. 13.

[48] P. Ekman, W.V. Friesen, The repertoire of nonverbal behavior: Categories, origins, usage, and coding, Semiotica 1 (1) (1969) 49–98.

[49] J. Akosa, Predictive accuracy: a misleading performance measure for highly imbalanced data, Proceedings of the SAS Global Forum (2017).

[50] Y. Sun, A.K. Wong, M.S. Kamel, Classification of imbalanced data: a review, Int. J. Pattern Recogn. Artif. Intell. 23 (4) (2009) 687–719.