Contents lists available at ScienceDirect

Expert Systems With Applications

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

Designing and evaluating a wearable device for affective state level classification using machine learning techniques

Luis Muñoz-Saavedra [a], Elena Escobar-Linero [b], Lourdes Miró-Amarante [c], M. Rocío Bohórquez [d], Manuel Domínguez-Morales [c,*]

[a] *Computer Engineering, Architecture and Computer Technology department (ATC), Robotics and Technology of Computers Lab (RTC), E.T.S. Ingeniería Informática, Universidad de Sevilla, Avda. Reina Mercedes s/n, Seville, 41012, Spain*
[b] *Biomedical Engineering, Architecture and Computer Technology department (ATC), Robotics and Technology of Computers Lab (RTC), E.T.S. Ingeniería Informática, Universidad de Sevilla, Avda. Reina Mercedes s/n, Seville, 41012, Spain*
[c] *Computer Engineering, Architecture and Computer Technology department (ATC), Robotics and Technology of Computers Lab (RTC), Computer Engineering Research Institute (I3US), E.T.S. Ingeniería Informática, Universidad de Sevilla, Avda. Reina Mercedes s/n, Seville, 41012, Spain*
[d] *Psychologist, Social Psychology department, Faculty of Psychology, Camilo José Cela s/n, Universidad de Sevilla, Seville, 41012, Spain*

ARTICLE INFO

Keywords:
Machine learning
Emotion
Affective state
Physiological signal
Wearable

ABSTRACT

The emotional or affective state has a direct impact not only on personal life, but also in the field of work, sports, rehabilitation processes, among other fields. In the evolving understanding of emotional theory, it has been theorized that an emotion can be classified according to a two-dimensional model composed of an *Arousal* value and a *Valence* value, as well as empirically demonstrating the impact of emotions on physiological variables. This work presents the development of a wearable device for capturing physiological signals, the collection of a dataset (after approval by the ethics committee) in which participants' emotional states are induced, and the development of an automatic classifier of the emotional state based on neural networks. According to this last point, a 4-phase optimization process is presented in which the physiological sensors are evaluated independently and with multiple variations of the hyperparameters of the neural networks, keeping those that provide the most information, combinations are made between them and the robustness of the final system obtained is evaluated. The results exceed 92% accuracy in all cases, which, compared with previous work, significantly improves the classifiers developed in recent years. The key contributions of this study are detailed as follows: (a) a wearable device designed to collect physiological signals from the user in a non-invasive way is presented, proving that it works properly in a controlled environment; (b) a data-collection protocol is designed to induce emotional states in test subjects using small video clips, demonstrating that the user evokes the feelings that are induced; and (c) a machine learning-based system is developed and optimized to classify the emotional state based on the two-dimensional model of emotion, demonstrating its efficiency and accuracy.

1. Introduction

Affective and emotional states have a direct impact on various aspects of daily life. One of the most affected aspects is work productivity: recent studies have revealed that workers' psychological states, such as positive attitudes and peace of mind, are associated with productivity at work. DiMaria, Peroni, and Sarracino (2020), Frey and Stutzer (2016) and Tenney, Poole, and Diener (2016) revealed that positive thinkers, i.e., happy and satisfied workers, are more likely to perform well. Studies have also highlighted how an employee's health and well-being impact productivity. Neumann and Dul (2010), as well as Ødegaard and Roos (2014), found that healthier employees are more productive.

Similarly, there are other areas in which a positive emotional state plays a key role, such as rehabilitation therapies (Hu, Xie, & Li, 2013; Te Wierike, van der Sluis, van den Akker-Scheek, Elferink-Gemser, & Visscher, 2013; Wiese-Bjornstal, Smith, Shaffer, & Morrey, 1998), or sports medicine (Ardern, Kvist, & Webster, 2016; Diener & Chan, 2011; Howell, Kern, & Lyubomirsky, 2007).

To combat the adverse effects that a negative emotional state can have, various control systems are implemented that make it possible to know the state of each worker and/or patient. In fact, some

[*] Correspondence to: E.T.S. Ingeniería Informática, Architecture and Computer Technology department (ATC), Avda. Reina Mercedes s/n, Office F0.71, 41012, Seville, Spain.

E-mail addresses: lmsaavedra@us.es (L. Muñoz-Saavedra), eescobar@us.es (E. Escobar-Linero), lmiro@us.es (L. Miró-Amarante), rociobohorquez@us.es (M.R. Bohórquez), mjdominguez@us.es (M. Domínguez-Morales).

https://doi.org/10.1016/j.eswa.2023.119577
Received 22 July 2022; Received in revised form 3 January 2023; Accepted 17 January 2023
Available online 20 January 2023

organizations have been increasing investment in health and wellness-related programs for years (Ton, 2014). In any case, these controls usually involve interviews with psychologists who, using classic metrics and surveys such as the *Discrete Emotions Questionnaire* (Harmon-Jones, Bastian, & Harmon-Jones, 2016), determine the patient's emotional state. However, these checks are not performed with a high cadence and, in most cases, are carried out when an adverse effect is detected (such as a significant reduction in productivity, or an unusually long recovery period for a particular injury).

In order to be able to detect any emotional signs that could cause problems in the future, it would be necessary to increase controls (without waiting to detect problems) or to develop a mechanism for automatic detection of the emotional state. For this purpose, multiple studies demonstrate the relationship between emotional state and various physiological parameters, mainly related to the unconscious reactions of the user's body that can be recorded and studied: facial muscle activity, cardiac activity, skin conductance, brain activity, among others.

Regarding mechanisms and/or tools that may help for automatic classification, the Artificial Intelligence (AI) field named "Machine Learning" (ML) has taken advantage in the last years. Some ML-tools like Support Vector Machines (SVM), Random Forest (RF), Logistic Regression (LR) or Neural Networks (NN) are commonly used actually for designing classifiers focused on the detection of diseases and/or anomalous medical states (Ayata, Yaslan, & Kamasak, 2020; Biagetti, Crippa, Falaschetti, Tanoni, & Turchetti, 2018; Kwon, Shin, & Kim, 2018; Santamaria-Granados, Munoz-Organero, Ramirez-Gonzalez, Abdulhay, & Arunkumar, 2018).

In multiple works, classifiers based on Machine Learning (ML) have been developed by this research group for anomaly detection using activity sensors (Domínguez-Morales, Luna-Perejón, Miró-Amarante, Hernández-Velázquez, & Sevillano-Ramos, 2019; Escobar-Linero, Domínguez-Morales, & Sevillano, 2022; Luna-Perejón, Domínguez-Morales, & Civit-Balcells, 2019; Luna-Perejón, Domínguez-Morales, Gutiérrez-Galán, & Civit-Balcells, 2020; Luna-Perejón, Muñoz-Saavedra, Civit-Masot, Civit, & Domínguez-Morales, 2021, or to detect diseases in medical images (Civit-Masot, Domínguez-Morales, Vicente-Díaz & Civit, 2020; Civit-Masot et al., 2021; Civit-Masot, Luna-Perejón, Domínguez Morales & Civit, 2020). Also, in a previous work, this team has even achieved acceptable emotional state classification results using a public dataset (Muñoz-Saavedra et al., 2020).

The main objectives of this work are the following: (a) designing and implementing a wearable device in order to capture the physiological data of the user, based on a previous work developed by this research group (Muñoz-Saavedra et al., 2020); (b) designing and testing a data collecting protocol from voluntary participants while inducing emotional states using audiovisual information; and (c) designing, implementing, optimizing and evaluating a Machine Learning-based classifier with a previous frequency features extraction using the data obtained from the empirical study.

This work is structured as follows: In the next section, a search of previous works related to emotional state classification using ML classifiers is carried out; in the third section, the tools and methodology used in this work (wearable device designing, data collecting protocol, and ML classifier designing) are presented; then, in the fourth section, the results of the process detailed in the previous section are shown, the system developed is compared with previous works, and the results obtained are discussed in detail. Finally, the conclusions obtained from this work are presented.

2. Related works

In this section, a search for similar works is performed in order to compare this work at the end of the manuscript. For this purpose, a global search is performed in the most commonly used search engines (IEEExplorer, ScienceDirect, and Google Scholar) using the following search sentence: ("emotional state" OR "affective state") AND ("deep learning" OR "machine learning") AND "physiological". The resulting set of works is filtered by year, restricting this parameter to those published from 2016 to 2022 (last seven years); and taking into account only those works published in international journals or congresses and only the most-cited works for each year. Preprints or arXiv/bioRxiv works waiting for acceptance are not selected.

The results after the search process are filtered by eliminating those that were not focused on a classifier design. The total number of works obtained is 17. The final selected works after the search process are briefly presented and summarized below, but their detailed results are included in the comparison table placed in the results section of this work:

- García, Álvarez, and Orozco (2016): in this work, a probabilistic dynamical model is performed on multimodal physiological signals related to affective state to classify between three classes. They use the DEAP dataset (32 participants induced with music videos) with EEG, EMG, EOG and GSR signals. The classifier used is a Supported Vector Machine (SVM).
- Liu, Meng, Nandi, and Li (2016): using only the EEG recordings from the same dataset as the previous work (DEAP), the authors use a k-Nearest neighbors and a Random Forest classifiers to distinguish between two classes.
- Li et al. (2016): using the DEAP dataset again and only the EEG signals, in this work a 2-class classifier using a combination of Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) is used.
- Zhang, Chen, Hu, Cao, and Kozma (2016): the DEAP dataset is used again with EEG signals too. Using a Probabilistic Neural Network (PNN), this work classifies between two classes.
- Mirmohamadsadeghi, Yazdani, and Vesin (2016): information regarding ECG and breathing patterns obtained from DEAP dataset, this work classifies two different classes using an SVM.
- Zheng, Zhu, and Lu (2017): for this work, three different classifiers are analyzed (k-NN, SVM and Logistic Regression) in order to distinguish between three different classes from DEAP dataset
- Girardi, Lanubile, and Novielli (2017): with the information given by DEAP dataset, a 2-class classifier based on a VSM is performed using EEG, GSR and EMG signals.
- Lee and Yoo (2018): in this case, a custom dataset is recorded using movie clips and bio-sensors (ECG, temperature and GSR). This dataset is used to develop a 2-class classifier using a neural network.
- Lee et al. (2019): in this work, authors use a convolutional neural network (CNN) for developing a 2-class classifier using DEAP dataset and the information provided by the photoplethysmograph.
- Sonkusare et al. (2019): a custom dataset is collected for this work, developing a CNN classifier for distinguishing two classes using ECG, GSR and temperature sensors.
- Lee and Yoo (2020): a classifier based on an RNN is developed for distinguishing between two classes using the information given by ECG, GSR and temperature. The dataset used are DEAP and EMDB.
- Domínguez-Jiménez, Campo-Landines, Martínez-Santos, Delahoz, and Contreras-Ortiz (2020): for recognizing between three different emotions (sad, joy and neutral), in this work an SVM classifier is developed using the information given by GRS and photoplethysmograph sensors from DEAP dataset.
- Ayata et al. (2020): DEAP dataset is used to train different classifiers based on RF, SVM and Logistic Regression to classify between two classes using the information given by the temperature, the breathing pattern and the photoplethysmograph.
- Sepúlveda, Castillo, Palma, and Rodriguez-Fernandez (2021): with the information given by the ECG signal from the AMIGOS dataset, the authors develop an ensemble bagged tree classifier to distinguish between two classes.

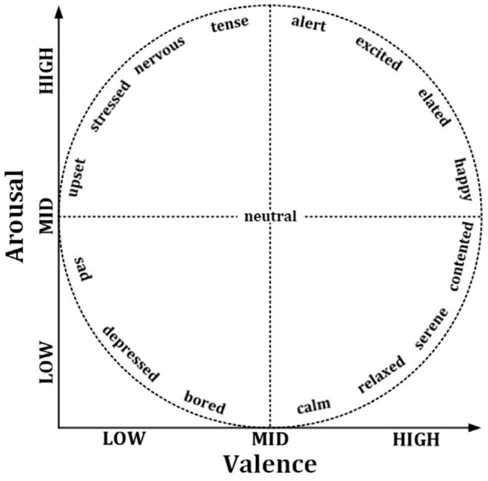


Fig. 1. Two-dimensional model based on *Valence* and *Arousal*.
Source: Adaptation from Mehrabian and Russel work.

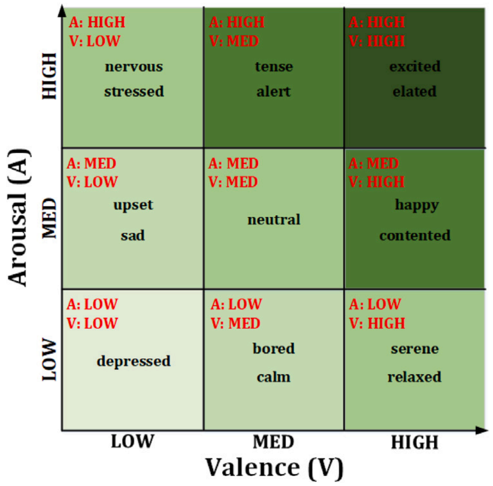


Fig. 2. Threshold division according to previous works (Muñoz-Saavedra et al., 2020).

- Tabbaa et al. (2021): In this work, authors develop a 2-class SVM classifier with the information collected in this same work (and shared as VREED dataset).
- Hammad and Monkaresi (2022): This author proposes a work where spatial and temporal features are extracted from the signals by a CNN, and later these features are used in a Deep Neural Network (DNN) in order to classify two classes. This work is tested with their own dataset and also with DREAMER dataset.
- Jalal and Peer (2022): Finally, in this work the authors propose a framework similar to our work, when they use the Continuous Wavelet Transform to extract information from the signals, and after that they obtained the features using a scalogram processing. The resulting features are used to train a CNN to distinguish between high and low values. They test their method with DEAP dataset.

In summary, the vast majority of the works make a classification of two classes (something that this work attempts to improve, increasing it to three classes to facilitate the recognition of emotions). Moreover, despite being a widely used mechanism, there are many papers that use classifiers not based on neural networks. And, finally, it can be observed that there is a little variety of emotion datasets, with DEAP and AMIGOS (by the same authors as DEAP) being the most widely used.

The following section will detail the emotional theory, the hardware device and collecting protocol used for this work and, finally, the optimizing process performed to obtain the classifier presented in this work. The comparison with the previous works' quantitative results is detailed in depth in the final part of the Results section.

3. Materials and methods

This section describes the tools used in this work, as well as the methodology used, starting with a quick explanation about the emotional theory.

3.1. Emotional theory

Throughout history, different emotions have been identified and different ways of classifying them have been proposed. The classification of emotions is based on a dimensional model of emotions, which allows a classification according to a series of values. These values have changed over the years.

But it is finally in 1974, when Mehrabian and Russel simplified the classification of emotions on two axes: *Valence* and *Arousal* (Mehrabian & Russell, 1974). Previous work has shown that this two-dimensional model of emotions is reliable for classifying emotions (see Fig. 1), and this model is the one used in this work too.

The dataset used in this work is labeled using the emotions induced by each case; and the emotions are labeled using three levels for *Valence* and *Arousal* (low, middle and high) according to Fig. 1. Additionally, we work with the thresholds used in a previous work (Muñoz-Saavedra et al., 2020) for this division, and include the labeling of the specific emotional states within these ranges according to Fig. 2. In summary, the numerical values labeled in public datasets were divided in that previous work into a range of the lowest 30% to define the "LOW" status, the highest 30% for the "HIGH" status and the middle 40% for the "MEDIUM" status.

3.2. Device and application

The hardware device designed for this work is composed of a STMicroelectronics microcontroller and two sensors for capturing physiological signals. The sensors that form it are:

- Galvanic Skin Response (GSR) sensor: it measures the electrogalvanic response of the skin, i.e. sweating. In order to measure this response, an electric current must be circulated through the body. This sensor uses two electrodes that are in contact with two fingers. Between these two electrodes there is an integrated circuit that measures the electrical conductivity; and the greater the sweating of the hand, the higher the value obtained.
- Photoplethysmograph sensor MAX30100: it consists of two light-emitting diodes, one emitting in the infrared spectrum (950 nm) and the other emitting red light (650 nm), and a photoreceptor. This sensor uses the absorption of blood at red and infrared wavelengths to measure both heart rate and blood oxygen saturation. To measure SPO_2 it is necessary to know the amount of red light absorbed by the blood, as well as the light absorbed in the infrared spectrum, the ratio between these two values will give us the percentage of oxygen carried by the blood. To measure the heart rate, the small changes that occur when measuring the maximum value of oxygen absorption in the blood will have to be measured, so the SPO_2 and heart rate values are calculated, and the values given by this module are the amount of light absorbed by the oxygenated hemoglobin (HbO_2), associated to the infrared spectrum and the amount of light absorbed by the deoxygenated hemoglobin (Hb), associated to the infrared spectrum. This sensor also includes an infrared thermometer.

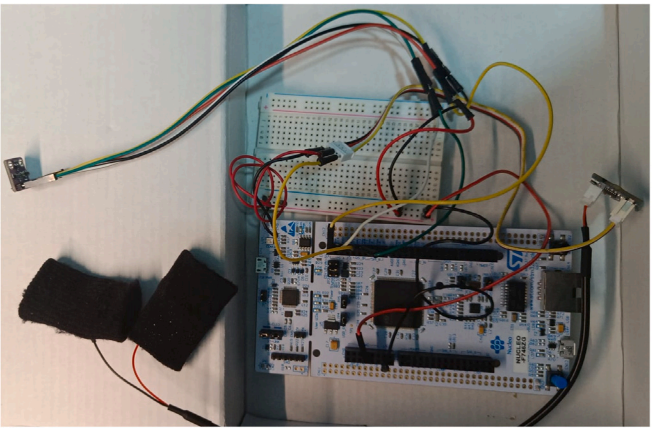


Fig. 3. Hardware device.

So, the physiological signals that are obtained with this device are:

- Sweating: obtained directly from the GSR sensor.
- Temperature: obtained directly from the MAX30100 sensor.
- Heart Rate: obtained indirectly from the HbO_2 signal of the MAX30100 sensor.
- Blood oxygen: obtained indirectly from the relation between Hb and HbO_2 signals of the MAX30100 sensor.

The prototype developed device can be observed in Fig. 3.

Regarding the data collecting software, it is integrated with the user application developed for the protocol performed. This application allows the user different tasks:

- User ID assignment: needed to distinguish the different participants after the collecting process.
- Connection with device: used for selecting the virtual COM port where the device is connected.
- Select path to store: selecting the folder where the data collected will be stored.
- Recording: there is a bidirectional communication between the user app and the device. The master of this communication is the app, which starts and stops the transmission using specific command. The information recorded is received at 50 hertz.
- Videoclip visualization: when the system starts collecting, due to the protocol (described next), the app stars showing some movie clips.

The main frame of the developed software is shown in Fig. 4.

After presenting the device and the user application developed for this work, the collection protocol followed and the resulting dataset will be detailed next.

3.3. *Dataset*

For the collection of the dataset used in this work, we have considered a population between 18 and 65 years of age, with basic handling of a digital device, who do not have vision difficulties and who do not suffer from any medical disorder that affects the recording of physiological samples (such as heart problems, lung problems, thyroid problems, nervous system disorders, etc.).

Following a protocol approved by the university's ethics committee, the device detailed above is used to collect physiological activity samples while inducing emotional states in the participants through a battery of videos (Megías, Mateos, Ribaudi, & Fernández-Abascal, 2011). The videos used in this process are the ones described in the previous work.

Table 1
Dataset samples collected for each class of *Arousal* and *Valence*.

		Temporal window (s)			
Class	Subset	2	3	4	5
LOW	Train	806	531	398	317
	Test	202	133	100	79
	Total	1008	664	498	396
MID	Train	299	187	147	114
	Test	75	49	37	28
	Total	374	246	184	142
HIGH	Train	605	403	299	235
	Test	151	101	75	59
	Total	756	504	374	294
GLOBAL	Train	1710	1131	845	666
	Test	428	283	211	166
	Total	2138	1414	1056	832

In addition, after each viewing, the user fills in the Discrete Emotions Questionnaire (DEQ), where he/she will determine what he/she feels after watching the video and what emotional state it has evoked in him/her. The whole process lasts between 45 and 60 minutes per participant.

The collection process is detailed in Fig. 5. In this figure it can be seen how, after viewing each video clip, the participant has an interview with an expert psychologist in which he/she is asked about the emotion evoked by the video he/she has watched, by filling in the DEQ. The result of this questionnaire is used as a label for the physiological signals collected during the video clip, taking into account that only those sections in which the corresponding event occurs in the video are labeled.

So, the physiological signals captured by the wearable device are subsequently tagged to an affective state (duple *Arousal* and *Valence*), taking into account the time stamps of the events in the videos.

Finally, after more than one month of collection, we managed to record information from 22 participants, 14 of whom were men and 8 women, with an average age of 31.67 years. Although the collected dataset does not contain numerous participants, it is large enough to be able to test the theory. It is also important to note that the dataset continues to expand week by week.

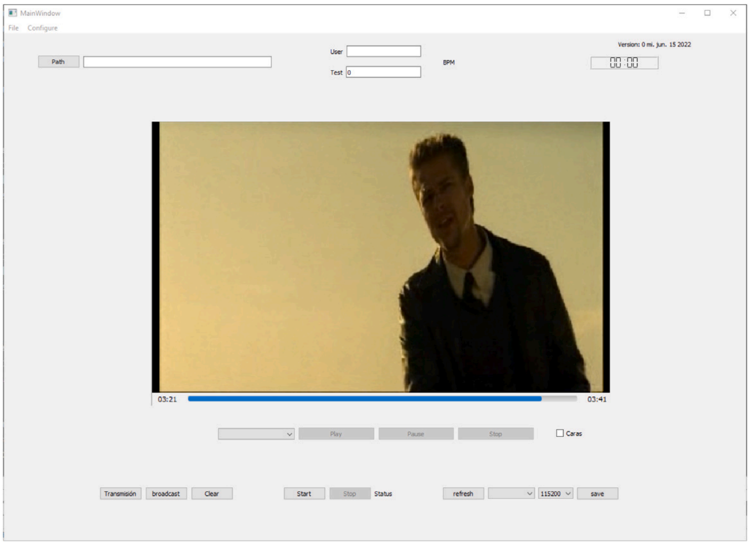
Since the duration of the videos varies from less than a minute to several minutes, numerous samples are obtained with the 22 participants since, as previously mentioned, the time windows to be analyzed vary between 2 and 5 seconds. The number of samples contained is the same for *Arousal* and *Valence*, and it is shown in Table 1.

It is important to note that the samples presented in Table 1 only include the number of temporal windows labeled for all the recordings (only a few seconds for each video). With the total amount of videos shown to each participant, the recording time collected for this test was more than 12 hours (and it was necessary more than 50 hours for this collection spread over six weeks, due to the pauses included between the videos and the notations and surveys performed for each participant).

3.4. *Classifier*

The classifier used for this work consists of a classical neural network (MLP, or multilayer perceptron) with an input layer, an output layer and two hidden layers.

The inputs to the classifier are the features extracted from the physiological signals of the participants, using a specific window width. Due to the nature of the sensors, these features may vary: specifically, the body temperature sensor provides an absolute value and, therefore, it is not useful to extract frequency features, but statistical variables; on the other hand, the sweating sensor and the photoplethysmograph (from which the oxygenated and deoxygenated hemoglobin signals



Version: 0 mi. jun. 15 2022

Test 0

115200

□ Caras

Fig. 4. User application.

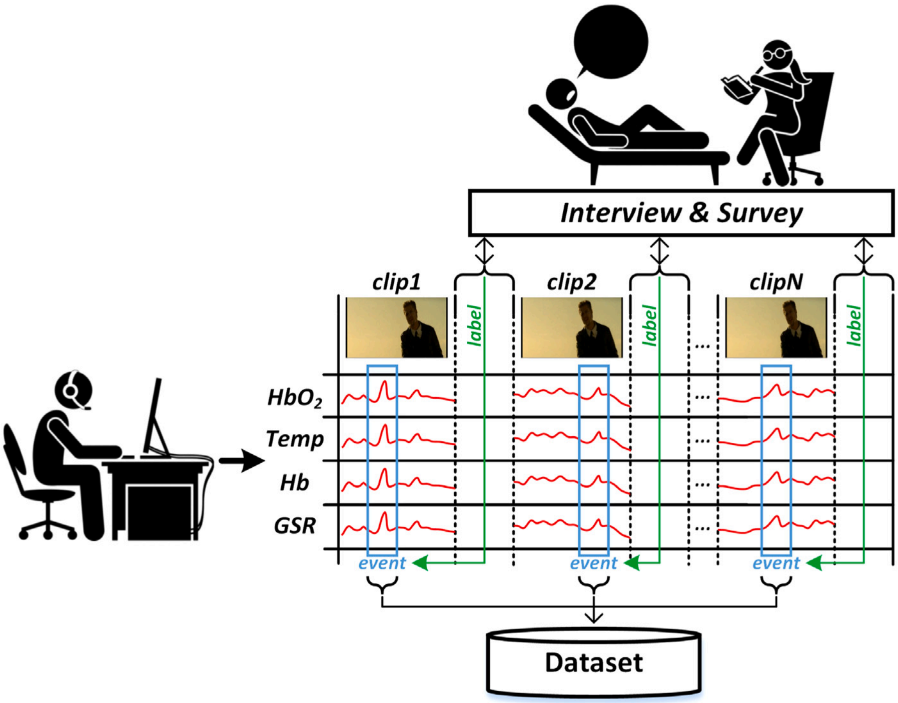


Fig. 5. Collection process.

are extracted), both provide oscillating and repetitive signals whose frequency information is of special interest.

In the case of this work, 5 temporal characteristics are extracted from the temperature sensor (arithmetic mean, quadratic mean, standard deviation, maximum and minimum), and 8 frequency characteristics from both the sweating signal and the oxygenated and deoxygenated hemoglobin signals (mean, deviation, amplitude and zero crossing of both the tonic and phasic signals). These features are summarized in Table 2.

Networks with different number of neurons and different values of hyperparameters (to be described later) are evaluated. In the same way, independent systems are trained for each sensor in order to, later on, realize combinations of sensors. This process can be summarized in Fig. 6.

The optimization process followed in this work to obtain the best classifier is divided into four phases that will be described below:

- Phase 1 — Hyperparameters adjustment: multiple trainings for each individual sensor are performed by varying the batch size and the time window values, as well as various alternatives of the architecture based on the number of layers and the number of neurons. In more detail, batch size is the parameter that indicates the number of samples used in each iteration of the training process before updating any parameter. As more samples are used in each iteration, the parameters will suffer less updates. However, very low samples used can lead to overfitting the model. Another parameter adjusted is the time window, which indicates the width (in seconds) of the window used to extract the frequency features. Other hyperparameters, like learning rate

Table 2
Features summary for each signal.

Signal	#	Features
Temperature	5	Min, Max, Arithmetic mean, Geometric mean, Standard deviation
HbO_2	8	Tonic: Zero crossings, Arithmetic mean, Standard deviation, Amplitude Phasic: Zero crossings, Arithmetic mean, Standard deviation, Amplitude
Hb	8	Tonic: Zero crossings, Arithmetic mean, Standard deviation, Amplitude Phasic: Zero crossings, Arithmetic mean, Standard deviation, Amplitude
GSR	8	Tonic: Zero crossings, Arithmetic mean, Standard deviation, Amplitude Phasic: Zero crossings, Arithmetic mean, Standard deviation, Amplitude

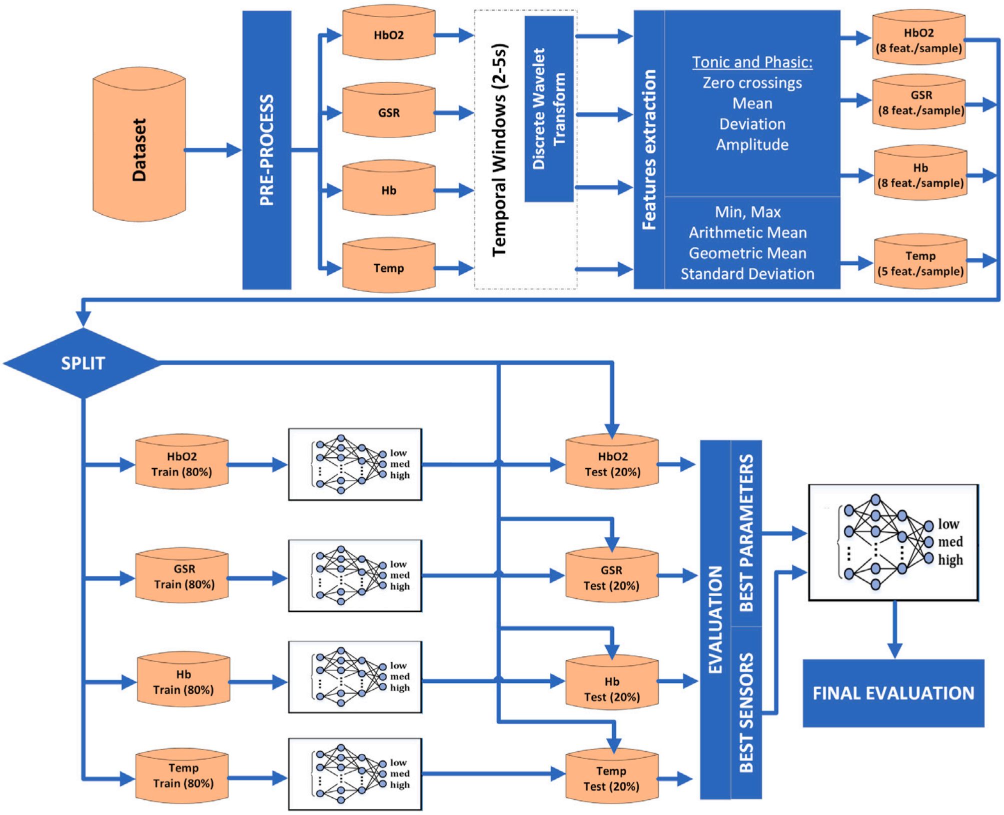


Fig. 6. Graphical abstract.

or dropout, are set following the results obtained in the previous work (Muñoz-Saavedra et al., 2020). Table 3 shows the summary of the hyperparameter values used in this grid search process. 36 combinations will be evaluated to obtain an optimal hyperparameter adjustment.

- Phase 2 — Best candidates selection: the previously obtained results are discussed and the best cases are selected, which will be used in the following phases.
- Phase 3 — Ensembles: the sensors that include the best candidates are joint forming an ensemble classifier with the information obtained from those sensors. Results are shown.
- Phase 4 — Exhaustive candidate assessment: more exhaustive tests are performed over the previously ensemble using different techniques to assess the robustness by using cross-validation, and the final results are detailed and discussed using different evaluation metrics.

Table 3
Hyperparameter's values.

Hyperparameter	Values
Learning rate	1e−4
Temporal window (s)	2, 3, 4, 5
Batch size	8, 16, 32
Hidden-layers size	32:16, 64:32, 128:64

3.5. Evaluation metrics

To evaluate the effectiveness in the classification results of a classifier, the most common metrics are used: accuracy (most-used metric), sensitivity (known as recall in other works), specificity, precision, and $F1_{score}$ (Sokolova et al., 2009). To this end, the classification results obtained for each class are tagged as "True Positive" (TP), "True Negative" (TN), "False Positive" (FP) or "False Negative" (FN). According to them, the high-level metrics are presented in the next equations:

$$\text{Accuracy} = \sum_{c} \frac{TP_c + TN_c}{TP_c + FP_c + TN_c + FN_c}, c \in classes \tag{1}$$

$$\text{Sensitivity} = \sum_{c} \frac{TP_c}{TP_c + FN_c}, c \in classes \tag{2}$$

$$\text{Specificity} = \sum_{c} \frac{TN_c}{TN_c + FP_c}, c \in classes \tag{3}$$

$$\text{Precision} = \sum_{c} \frac{TP_c}{TP_c + FP_c}, c \in classes \tag{4}$$

$$F1_{score} = 2 * \frac{\text{precision} * \text{sensitivity}}{\text{precision} + \text{sensitivity}}. \tag{5}$$

Table 4
Results obtained for HbO_2 signal for the three architectures when classifying *Arousal*.

Architecture	Window (s)	Batch size											
		8				16				32			
		Train		Test		Train		Test		Train		Test	
		Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss
128:64	2	96.82	0.0964	99.66	0.0609	96.76	0.0899	93.42	0.7574	94.63	0.1355	95.01	0.2113
	3	98.38	0.0524	100.00	0.0150	98.45	0.0433	65.10	94.3020	42.73	1.0775	62.23	0.9836
	4	99.42	0.0174	100.00	0.0089	98.18	0.0569	100.00	0.0377	98.56	0.0564	100.00	0.0335
	5	99.20	0.0368	75.44	4.1654	97.87	0.0634	100.00	0.0496	99.02	0.0393	91.30	3.0769
64:32	2	86.06	0.3660	83.77	0.4447	85.37	0.3532	77.23	0.7721	87.74	0.3270	70.31	1.1107
	3	92.52	0.2106	97.93	0.1256	90.83	0.2544	93.75	0.2224	89.66	0.2773	87.32	0.3727
	4	93.82	0.1615	98.00	0.0780	91.67	0.2276	89.38	0.3155	93.05	0.1894	93.20	0.3107
	5	91.67	0.2533	92.22	0.3457	92.24	0.2187	97.75	0.1034	90.87	0.2394	78.23	1.1948
32:16	2	40.02	1.3512	80.66	0.7006	75.16	0.5479	57.12	1.1362	41.00	1.1200	57.19	1.0101
	3	77.61	0.5854	74.31	0.8201	23.78	1.6896	51.97	1.0819	74.45	0.5789	69.61	0.7072
	4	85.22	0.4142	65.00	1.5811	80.67	0.4484	72.43	0.9505	73.93	0.5892	74.11	0.7346
	5	86.37	0.3616	77.02	0.7553	83.58	0.4027	58.85	52.6437	80.74	0.4558	79.82	0.4815

Table 5
Results obtained for HbO_2 signal for the three architectures when classifying *Valence*.

Architecture	Window (s)	Batch size											
		8				16				32			
		Train		Test		Train		Test		Train		Test	
		Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss
128:64	2	97.76	0.0669	76.89	3.1621	96.64	0.1056	99.54	0.0603	94.85	0.1435	95.76	0.1545
	3	99.50	0.0263	93.20	0.7699	97.42	0.0923	98.68	0.0714	98.15	0.0536	95.20	0.5982
	4	64.16	0.7936	77.97	0.9352	99.23	0.0243	97.30	0.4014	96.55	0.1161	99.68	0.0659
	5	99.27	0.0343	97.48	0.1278	98.87	0.0341	88.71	2.1190	98.04	0.0799	54.72	5.4271
64:32	2	88.09	0.3178	75.85	0.8349	88.88	0.2874	65.03	2.9453	85.31	0.3849	81.16	0.4762
	3	91.74	0.2296	89.96	0.7649	88.80	0.2921	89.56	0.4209	87.26	0.3445	87.17	0.3021
	4	95.54	0.1333	100.00	0.0459	94.00	0.1857	95.37	0.2365	92.48	0.1903	89.78	0.5080
	5	95.05	0.1467	98.47	0.1198	95.76	0.1419	94.95	0.2163	88.08	0.2962	98.39	0.1497
32:16	2	70.95	0.6808	73.08	0.6246	74.35	0.5782	46.32	1.1140	76.67	0.5274	72.30	0.6236
	3	82.09	0.4363	74.31	0.6553	75.28	0.5367	77.60	0.5379	75.89	0.5597	68.20	0.6840
	4	82.22	0.4517	79.64	0.4986	81.84	0.4340	73.52	0.9603	74.83	0.6006	70.86	0.7179
	5	87.94	0.3089	66.53	1.4393	83.17	0.4508	73.09	0.6896	78.68	0.4862	84.00	0.4635

About those metrics:

- Accuracy: all samples classified correctly compared to all samples (see Eq. (1)).
- Sensitivity (or recall): proportion of values classified as "true positive" that are correctly classified (see Eq. (2)).
- Specificity: proportion of values classified as "true negative" that are correctly classified (see Eq. (3)).
- Precision: proportion of values classified as "true positive" in all cases that have been classified as it (see Eq. (4)).
- $F1_{score}$: It considers two of the main metrics (precision and sensitivity), calculating the harmonic mean of both parameters (see Eq. (5)).

The above metrics are common to all ML systems; but there are other commonly used metrics in healthcare systems; this is the case of the ROC curve (Receiver Operating Characteristic) (Hoo, Candlish, & Teare, 2017), because it is the visual representation of the True Positives Rate (TPR) versus the False Positives Rate (FPR) as the discrimination threshold is varied. Usually, when using the ROC curve, the area under the curve (AUC) is used as a value of the system's goodness-of-fit. Therefore, the classifier system developed in this work will be evaluated according to all the metrics detailed in this subsection.

4. Results and discussions

The results obtained in the four phases used to optimize the classifier will be presented in order. Similarly, as results are presented, the candidates chosen in each phase will be determined according to the evaluation metrics obtained.

Once the best classifier is obtained, a thorough comparison will be made with the works detailed in Section 2.

4.1. Classifier implementation and evaluation

Starting with the design and evaluation of the classifier, independent classifiers will be trained for each sensor, the best candidates will be extracted and, based on them, combinations will be made to obtain the final results.

4.1.1. Phase 1: Grid search

The results in this section will be presented as follows: for each signal, two tables will be shown: one for *Arousal* and the other for *Valence*. Each of these tables will contain the classification results (accuracy and loss) for the training and testing subsets for each combination of hyperparameter (batch size: 8, 16 and 32; and temporal window: 2, 3, 4 and 5) and for each network architecture (128:64, 64:32 and 32:16).

First, results for HbO_2 signal are shown in Table 4 for *Arousal*, and in Table 5 for *Valence*.

As can be observed in Table 4, the best results seem to be located in the temporal width of four seconds for the most complex architecture (128:64). Acceptable results are colored in yellow, good results in green and bad results in red. Even so, it can be seen how this signal with this architecture achieves good classification results. Moreover, as presented in Table 5, there are some good results with particular parameters, but it seems that this signal has more difficulty to classify *Valence*.

Table 6
Results obtained for temperature signal for the three architectures when classifying *Arousal*.

Architecture	Window (s)	Batch size											
		8				16				32			
		Train		Test		Train		Test		Train		Test	
		Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss
128:64	2	76.95	0.5172	67.97	0.6919	77.30	0.5124	66.49	0.7339	73.45	0.5998	84.55	0.4458
	3	78.72	0.4849	74.28	0.6006	74.71	0.5726	79.69	0.6837	77.35	0.5291	69.01	0.7845
	4	72.80	0.6069	87.76	0.4371	75.11	0.5356	81.95	0.8057	79.13	0.4963	76.43	0.6124
	5	76.41	0.5247	74.17	0.8297	80.18	0.4552	62.30	0.8469	78.00	0.4943	89.09	0.4082
64:32	2	67.92	0.7016	70.52	0.7744	74.17	0.5891	57.00	3.1390	69.82	0.6394	84.95	0.4721
	3	60.87	0.8337	69.29	0.8742	75.36	0.5977	60.41	0.8467	73.28	0.5812	67.42	7.1094
	4	75.64	0.5494	68.77	0.8640	74.05	0.5907	80.14	0.5523	74.16	0.5965	80.33	0.5245
	5	48.56	1.0043	67.41	0.9441	70.73	0.6474	59.89	1.1397	73.52	0.6286	72.16	0.7214
32:16	2	59.72	0.8848	49.36	1.0804	54.92	0.9303	52.09	1.1098	68.22	0.6724	72.37	0.7119
	3	62.80	0.8037	53.72	1.0854	67.18	0.7144	84.55	0.4601	61.26	0.8315	65.17	0.8217
	4	72.90	0.6409	61.54	0.8814	68.73	0.7108	69.43	1.1181	35.55	3.4068	72.79	0.8866
	5	66.86	0.7537	74.19	0.7535	73.58	0.6144	75.60	0.6375	67.63	0.7383	68.90	0.8337

Table 7
Results obtained for temperature signal for the three architectures when classifying *Valence*.

Architecture	Window (s)	Batch size											
		8				16				32			
		Train		Test		Train		Test		Train		Test	
		Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss
128:64	2	73.59	0.5765	59.95	0.9190	70.31	0.6622	62.74	0.9986	72.18	0.6165	87.28	0.3796
	3	71.53	0.6191	81.01	0.5996	74.69	0.5670	74.79	0.6597	74.40	0.6111	72.88	0.6619
	4	76.59	0.5474	77.70	0.5398	72.88	0.6434	53.81	2.9568	73.33	0.6337	70.51	0.7008
	5	78.81	0.4882	72.13	0.8003	78.76	0.4555	77.10	0.5534	79.45	0.4718	82.01	0.5040
64:32	2	71.77	0.6190	60.99	0.9361	69.99	0.7037	56.11	0.9875	74.14	0.5815	70.92	0.6701
	3	72.53	0.6455	72.92	0.6522	71.58	0.6407	55.84	1.0169	75.24	0.5731	69.48	0.8485
	4	76.94	0.5242	64.57	0.9307	71.11	0.6475	70.28	0.8665	65.66	0.7527	68.34	0.8473
	5	69.73	0.6372	82.69	0.5349	67.70	0.7176	67.29	0.7782	74.75	0.5728	56.30	1.3354
32:16	2	43.80	1.0522	67.24	1.0676	70.13	0.6706	71.65	0.7320	65.17	0.7606	66.01	0.8367
	3	71.46	0.6284	70.37	0.7747	67.77	0.7357	59.72	1.0666	67.16	0.6960	61.95	0.9288
	4	63.07	0.8316	71.57	0.7452	71.34	0.6408	73.73	0.8750	64.45	0.7808	74.57	0.7612
	5	59.34	0.8587	64.95	0.8245	72.87	0.6514	62.03	1.0203	61.74	0.8240	52.75	1.1462

For the 64:32 architecture, classifiers' accuracy decreased due to the architecture complexity reduction, however some cases obtain acceptable results (over 90%). Finally, as expected, results are clearly worse with the 32:16 architecture than the ones obtained in the previous ones. However, accuracy stands around 85% in some cases for *Arousal* and *Valence*.

Secondly, results for temperature signal are shown in Table 6 for *Arousal*, and in Table 7 for *Valence*.

As can be observed in Table 6, the temperature sensor rating results for *Arousal* are mostly poor for the most complex architecture. Therefore, it is clear that this sensor would not be suitable in this case. Moreover, the temperature sensor does not perform well when classifying *Valence*. The best cases obtain around 80%–89% accuracy for *Arousal* and even less for *Valence*.

For the 64:32 architecture, the results are even worse: The best cases obtain around 80%–84% accuracy. And, finally, for the lighter architecture, accuracy results obtained are around 60%–70%.

In third place, results for GSR signal are shown in Table 8 for *Arousal*, and in Table 9 for *Valence*.

As can be observed in Table 8, the classification results for the GSR signal with *Arousal* are not very good. There are cases where acceptable results are obtained with the 128:64 architecture, but the vast majority are low results. For *Valence*, the GSR signal seems to obtain acceptable classification results with the heavier architecture, exceeding 94% in most cases. Summarizing, GSR sensor may be used for *Valence* classification, but next results will show if there is no other sensor that obtains better results.

After some promising results with the 128:64 architecture, the middle architecture's results continues the same tendency by obtaining very good results (over 90% in most cases). And, finally, for the lighter architecture, although some results are not acceptable, in the most cases results over 85% are obtained. These results give this sensor the first position in terms of accuracy so far.

The fourth signal evaluated if the deoxygenated hemoglobin (Hb). Its results for Hb signal are shown in Table 10 for *Arousal*, and in Table 11 for *Valence*.

Results in Table 10 shown that there are acceptable results for classifying *Arousal* with Hb signal using the heavier architecture, although *Loss* values are too high. Results for *Valence* show that acceptable results are obtained for Hb signal, but there are several values with high *Loss* values.

Although, with the 128:64 architecture, results were very promising, with the 64:32 architecture the accuracy plummet this time. There are some cases where accuracy surrounds 90%, but most of them are around 80%–85%, and worse results are obtained for *Arousal*. Finally, for the lighter architecture, as happened before, results obtained are lower. In comparison with the other signals, it obtains worse results than HbO_2 and GSR sensors, but not as low as temperature sensor.

The results presented are summarized as follows:

- Regarding the results obtained exclusively for the 128:64 architecture, two signals stand out: HbO_2 and sweating (GSR sensor). Temperature sensor is the one with the worst results, and Hb signal presents too high *Loss* values.
- Regarding the results with the other architectures, some conclusions could be made: temperature sensor is not suitable to be used, and Hb signal experiments a high decreasing from the first architecture to this one. So, HbO_2 and GSR seems to be the signals that provide the most information.

Table 8
Results obtained for GSR signal for the three architectures when classifying *Arousal*.

Architecture	Window (s)	Batch size											
		8				16				32			
		Train		Test		Train		Test		Train		Test	
		Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss
128:64	2	98.90	0.0368	80.58	4.2113	97.53	0.0711	74.23	4.6345	96.97	0.0795	72.62	11.3095
	3	97.11	0.0787	99.33	0.0538	94.62	0.1527	65.98	6.7071	96.45	0.1113	99.02	0.0912
	4	97.75	0.0706	100.00	0.0401	95.98	0.1020	100.00	0.0347	92.07	0.2168	90.91	0.3499
	5	96.20	0.0779	99.12	0.0556	97.45	0.0721	81.61	3.0394	93.84	0.1424	92.80	0.4396
64:32	2	93.29	0.1840	94.86	0.1909	91.37	0.2325	89.35	0.3223	93.76	0.1744	70.72	1.9500
	3	89.39	0.2680	92.05	0.2439	90.49	0.2840	92.37	0.5006	84.16	0.3977	90.71	0.2729
	4	92.27	0.1972	78.76	0.7776	90.66	0.2498	92.22	0.4018	87.44	0.3114	92.69	0.3024
	5	92.82	0.2053	96.05	0.1896	91.04	0.2189	89.15	0.4125	85.42	0.3828	88.56	0.3544
32:16	2	84.02	0.4255	87.33	0.3317	83.39	0.4428	82.56	0.4626	76.54	0.6018	69.82	0.8183
	3	87.14	0.3266	86.85	0.3478	82.29	0.4534	67.28	1.3948	82.56	0.4612	65.79	0.7522
	4	85.35	0.3774	79.55	0.5427	83.31	0.4070	80.31	0.5278	79.33	0.5642	62.15	1.0934
	5	84.72	0.4255	82.07	0.5128	75.24	0.6139	75.69	0.6652	77.06	0.5643	75.76	0.6351

Table 9
Results obtained for GSR signal for the three architectures when classifying *Valence*.

Architecture	Window (s)	Batch size											
		8				16				32			
		Train		Test		Train		Test		Train		Test	
		Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss
128:64	2	96.66	0.1079	100.00	0.0228	97.75	0.0641	100.00	0.0298	98.12	0.0539	100.00	0.0076
	3	96.11	0.1018	100.00	0.0404	97.32	0.0773	100.00	0.0391	95.32	0.1373	100.00	0.0489
	4	94.31	0.1522	100.00	0.0454	96.73	0.0858	95.96	2.3339	94.27	0.1246	100.00	0.0495
	5	96.48	0.0932	75.00	7.4356	94.60	0.1398	68.71	10.5939	95.56	0.1158	89.17	0.8851
64:32	2	94.22	0.1612	97.78	0.0715	91.47	0.2465	94.42	0.2036	91.00	0.2455	87.06	0.4023
	3	93.05	0.1812	97.47	0.1212	90.94	0.2261	75.94	6.6802	85.92	0.3749	86.83	0.3579
	4	93.42	0.2030	94.74	0.1974	89.71	0.2897	75.00	2.2000	88.55	0.3019	91.39	0.3123
	5	91.58	0.2197	93.89	0.1981	86.35	0.3734	85.09	1.2349	87.35	0.3415	63.81	4.3691
32:16	2	71.63	0.6483	81.98	0.6813	85.15	0.3947	84.80	0.4549	40.60	2.3906	60.47	0.9503
	3	83.89	0.3918	70.50	0.9630	80.41	0.5014	89.77	0.4212	86.33	0.3597	57.45	4.6116
	4	85.51	0.3623	66.18	1.0666	47.10	1.0689	68.75	1.0017	77.52	0.5355	75.54	0.6656
	5	48.62	1.1818	73.15	1.0565	81.32	0.4861	63.80	3.1625	79.37	0.5235	80.98	0.5281

Table 10
Results obtained for Hb signal for the three architectures when classifying *Arousal*.

Architecture	Window (s)	Batch size											
		8				16				32			
		Train		Test		Train		Test		Train		Test	
		Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss
128:64	2	93.33	0.1752	97.27	0.1558	92.36	0.2106	95.58	0.1473	92.75	0.1938	92.53	0.2251
	3	98.19	0.0511	66.53	4.2423	95.28	0.1421	96.69	0.1441	92.02	0.2109	89.25	0.7027
	4	97.24	0.0733	87.60	6.8105	95.31	0.1245	66.12	6.5035	96.45	0.1089	83.92	1.7006
	5	97.90	0.0730	99.53	0.0606	97.38	0.0617	76.99	3.4677	93.46	0.1698	80.08	1.5687
64:32	2	77.75	0.5148	77.55	0.5431	82.70	0.4061	80.43	0.4695	75.80	0.5554	77.12	0.7400
	3	88.71	0.2999	91.05	0.2867	85.22	0.3742	90.07	0.3927	84.96	0.3967	86.78	0.7335
	4	85.66	0.3368	84.08	0.9400	91.13	0.2647	81.82	0.7712	84.47	0.4056	80.20	0.5402
	5	90.20	0.2706	72.86	2.7277	89.80	0.3212	89.45	0.3221	90.69	0.2457	51.16	5.8066
32:16	2	62.05	0.7986	60.42	0.9100	65.90	0.7420	75.38	0.6226	48.42	1.0092	69.16	0.9327
	3	72.54	0.6568	60.96	0.8193	76.11	0.5575	58.85	1.3655	70.59	0.6886	69.58	0.8468
	4	77.69	0.5409	58.90	0.8819	67.35	0.7183	72.92	0.7040	73.65	0.6118	47.87	1.1607
	5	79.51	0.4971	55.94	1.0931	73.72	0.5942	74.23	0.7133	72.87	0.6297	74.24	0.6313

Therefore, concluding a first reflection on the results obtained individually by each sensor, the temperature sensor would be the first to be discarded, while the Hb sensor would have to be compared more closely with the other two. In the following subsection, the best results obtained by each of these three sensors (definitely discarding the temperature sensor) will be summarized and the best candidates to be used in the combination phase will be discussed.

4.1.2. *Phase 2: Best individual candidates*

The objective of this section is to present, in a more summarized form, the information previously presented. Due to the large amount of information previously shown, it is not easy to simplify it much (so results per sensor and architecture will be presented equally), but the worst sensor (temperature) and the worst results of the remaining ones will be eliminated. In summary, one table per sensor will be presented.

Table 11
Results obtained for Hb signal for the three architectures when classifying *Valence*.

Architecture	Window (s)	Batch size											
		8				16				32			
		Train		Test		Train		Test		Train		Test	
		Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss
128:64	2	94.55	0.1385	82.71	1.3346	50.83	0.9680	65.70	0.9635	91.50	0.2319	83.33	1.6608
	3	95.55	0.1304	96.53	0.2801	95.53	0.1282	96.20	1.0421	95.05	0.1489	90.08	1.4373
	4	97.91	0.0583	100.00	0.0198	93.92	0.1679	61.90	2.7557	95.54	0.1286	95.24	1.6512
	5	98.55	0.0414	63.64	9.1993	98.32	0.0703	87.10	9.3770	96.43	0.1163	90.26	0.8020
64:32	2	79.05	0.5137	70.60	0.7818	77.80	0.5096	76.83	0.6720	78.89	0.4862	64.95	2.0814
	3	85.94	0.3626	50.71	4.3060	85.25	0.3994	67.81	1.0739	85.49	0.3708	77.99	0.5586
	4	88.27	0.2922	89.08	0.3616	85.05	0.3653	79.53	0.6276	81.07	0.4399	73.26	0.9680
	5	89.52	0.2756	92.05	0.2565	87.69	0.3096	86.60	0.3688	90.46	0.2434	60.87	2.2712
32:16	2	69.77	0.6741	65.43	0.7743	67.95	0.7249	37.33	1.2434	74.03	0.6133	44.61	1.6463
	3	76.22	0.5568	46.37	2.3196	72.65	0.6053	71.69	1.0892	72.74	0.6415	54.87	0.9327
	4	80.64	0.4429	61.87	2.3723	73.62	0.6192	67.70	0.6946	71.55	0.5960	70.56	0.6391
	5	71.68	0.5609	52.96	1.9513	73.21	0.6024	71.76	0.7228	58.60	0.8745	63.69	1.1459

Table 12
Best results regarding the three architectures for HbO_2 signal.

Architecture	Window (s)	Best HbO_2 results									
		Arousal					Valence				
			Train		Test			Train		Test	
		BS	Acc (%)	Loss	Acc (%)	Loss	BS	Acc (%)	Loss	Acc (%)	Loss
128:64	2	8	96.82	0.0964	99.66	0.0609	16	96.64	0.1056	99.54	0.0603
	3	8	98.38	0.0524	100.00	0.0150	16	97.42	0.0923	98.68	0.0714
	4	8	99.42	0.0174	100.00	0.0089	32	96.55	0.1161	99.68	0.0659
	5	16	97.87	0.0634	100.00	0.0496	8	99.27	0.0343	97.48	0.1278
64:32	2	8	86.06	0.3660	83.77	0.4447	32	85.31	0.3849	81.16	0.4762
	3	8	92.52	0.2106	97.93	0.1256	8	91.74	0.2296	89.96	0.7649
	4	8	93.82	0.1615	98.00	0.0780	8	95.54	0.1333	100.00	0.0459
	5	16	92.24	0.2187	97.75	0.1034	8	95.05	0.1467	98.47	0.1198
32:16	2	8	70.95	0.6808	73.08	0.6246	8	40.02	1.3512	80.66	0.7006
	3	16	75.28	0.5367	77.60	0.5379	8	77.61	0.5854	74.31	0.8201
	4	8	82.22	0.4517	79.64	0.4986	32	73.93	0.5892	74.11	0.7346
	5	32	78.68	0.4862	84.00	0.4635	32	80.74	0.4558	79.82	0.4815

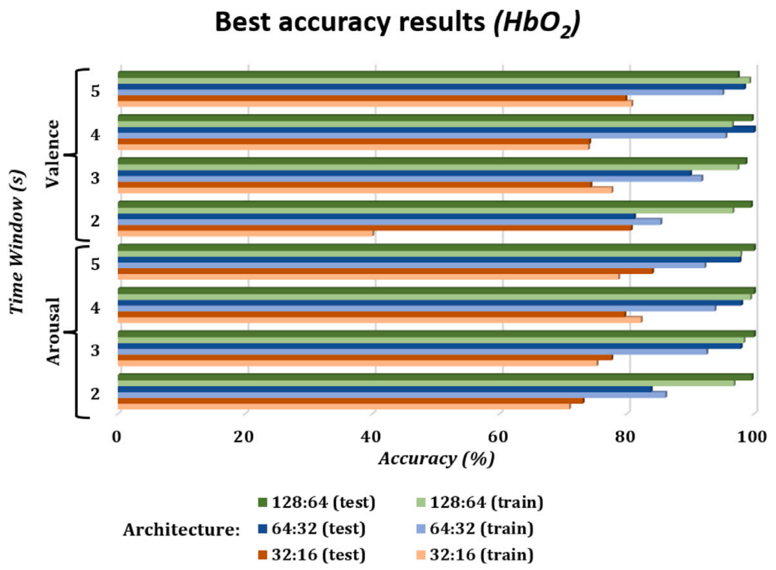


Fig. 7. Graphic representation of the best results obtained for HbO_2 signal.

After the best results are presented, we will be able of decide which sensors are most suitable for the final classifier design.

Table 12 presents the best results obtained for HbO_2 signal with the three architectures. For the 128:64 architecture, although better results are obtained for *Arousal*, all cases surpass 97% accuracy. With the second architecture (64:32), sensor HbO_2 decreases slightly the accuracy, but results remain acceptable with values over 98% for both *Arousal* and *Valence*. Finally, for this sensor, results for the 32:16 architecture are clearly worse, but some results are around 80% accuracy. Although,

Table 13
Best results regarding the three architectures for GSR signal.

Architecture	Window (s)	Best GSR results									
		Arousal					Valence				
			Train		Test			Train		Test	
		BS	Acc (%)	Loss	Acc (%)	Loss	BS	Acc (%)	Loss	Acc (%)	Loss
128:64	2	8	98.90	0.0368	80.58	4.2113	8	96.66	0.1079	100.00	0.0228
	3	8	97.11	0.0787	99.33	0.0538	8	96.11	0.1018	100.00	0.0404
	4	8	97.75	0.0706	100.00	0.0401	8	94.31	0.1522	100.00	0.0454
	5	8	96.20	0.0779	99.12	0.0556	32	95.56	0.1158	89.17	0.8851
64:32	2	8	93.29	0.1840	94.86	0.1909	8	94.22	0.1612	97.78	0.0715
	3	16	90.49	0.2840	92.37	0.5006	8	93.05	0.1812	97.47	0.1212
	4	32	87.44	0.3114	92.69	0.3024	8	93.42	0.2030	94.74	0.1974
	5	8	92.82	0.2053	96.05	0.1896	8	91.58	0.2197	93.89	0.1981
32:16	2	8	84.02	0.4255	87.33	0.3317	16	85.15	0.3947	84.80	0.4549
	3	8	87.14	0.3266	86.85	0.3478	16	80.41	0.5014	89.77	0.4212
	4	16	83.31	0.4070	80.31	0.5278	32	77.52	0.5355	75.54	0.6656
	5	8	84.72	0.4255	82.07	0.5128	32	79.37	0.5235	80.98	0.5281

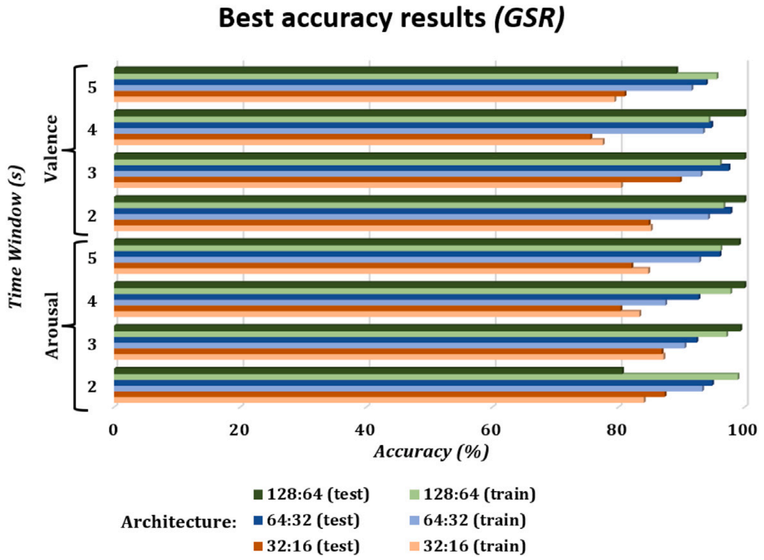


Fig. 8. Graphic representation of the best results obtained for GSR signal.

individually, the lighter architecture does not obtain very high results, it may be improved by the combination with other sensors, so it is not discarded yet. These observations can also be seen in Fig. 7.

Secondly, Table 13 presents the best results obtained for GSR signal with the three architectures. In this case, better results are obtained for *Valence*, but almost all cases surpass 90% accuracy. With the second architecture, GSR sensor decreases slightly the accuracy, but results remain acceptable with values over 92% for both *Arousal* and *Valence* (with better results for *Valence*). Finally, for GSR sensor, results for the lightest architecture decrease significantly, but they are over 80% for all cases, and even some results around 90% are obtained. So far, this sensor seems to be the one with the best results for all architectures. These observations can also be seen in Fig. 8.

Finally, Table 14 presents the best results obtained for Hb signal with the three architectures. It can be observed that good results are achieved (over 85% for most cases), but they are worse than the ones obtained with the other sensors. With the second architecture, Hb signal decreases the accuracy: In some cases this decreasing is low (around 10%), but in other cases the results decrease more than a 20%. And finally, for Hb signal, accuracy results obtained for the lightest architecture are getting worryingly worse, reducing the accuracy to values around 70% in most cases. These observations can also be seen in Fig. 9.

After presenting the best results, we can draw an important conclusion: the signal whose accuracy is most significantly reduced by the change of architecture is Hb. The other two obtain worse results with less heavy architectures, but they are still acceptable results (above 80% in most cases).

In a previous work carried out by this research group (Muñoz-Saavedra et al., 2020), the conclusions obtained are similar: the two physiological signals that provide the most information for the classification of emotional state are sweating and heart rate. It is true that the HbO_2 signal is not exactly the same as an ECG (as used in that previous study) but, by using frequency features extracted from the DWT, the information obtained indirectly are the peaks of the signal which, effectively, are correlated with the heart rate. So, it is clear that the two signals used for the ensemble network will be GSR and HbO_2.

In that previous work, the best signal for *Arousal* and *Valence* classifications was GSR with a result between 82 and 83% accuracy; on the other hand, the second-best signal was ECG with a result around 79%–80%. These results are really similar that the ones obtained with the lighter architecture of this work. Moreover, the networks used in the previous work were two hidden layers with 24 to 32 neurons for the first hidden layer, and with 6 to 12 neurons for the second hidden layer.

As we want to obtain a light classifier in order to be integrated in an embedded system, the sensor combination used in next subsection will use the lighter architecture (32 neurons for the first hidden layer and 16 for the second one). It is true that the results of this architecture

Table 14
Best results regarding the three architectures for Hb signal.

Architecture	Window (s)	Best Hb results									
		Arousal					Valence				
			Train		Test			Train		Test	
		BS	Acc (%)	Loss	Acc (%)	Loss	BS	Acc (%)	Loss	Acc (%)	Loss
128:64	2	8	93.33	0.1752	97.27	0.1558	32	91.50	0.2319	83.33	1.6608
	3	16	95.28	0.1421	96.69	0.1441	8	95.55	0.1304	96.53	0.2801
	4	8	97.24	0.0733	87.60	6.8105	8	97.91	0.0583	100.00	0.0198
	5	8	97.90	0.0730	99.53	0.0606	32	96.43	0.1163	90.26	0.8020
64:32	2	16	82.70	0.4061	80.43	0.4695	16	77.80	0.5096	76.83	0.6720
	3	8	88.71	0.2999	91.05	0.2867	32	85.49	0.3708	77.99	0.5586
	4	8	85.66	0.3368	84.08	0.9400	8	88.27	0.2922	89.08	0.3616
	5	16	89.80	0.3212	89.45	0.3221	8	89.52	0.2756	92.05	0.2565
32:16	2	16	65.90	0.7420	75.38	0.6226	8	69.77	0.6741	65.43	0.7743
	3	32	70.59	0.6886	69.58	0.8468	16	72.65	0.6053	71.69	1.0892
	4	16	67.35	0.7183	72.92	0.7040	32	71.55	0.5960	70.56	0.6391
	5	32	72.87	0.6297	74.24	0.6313	16	73.21	0.6024	71.76	0.7228

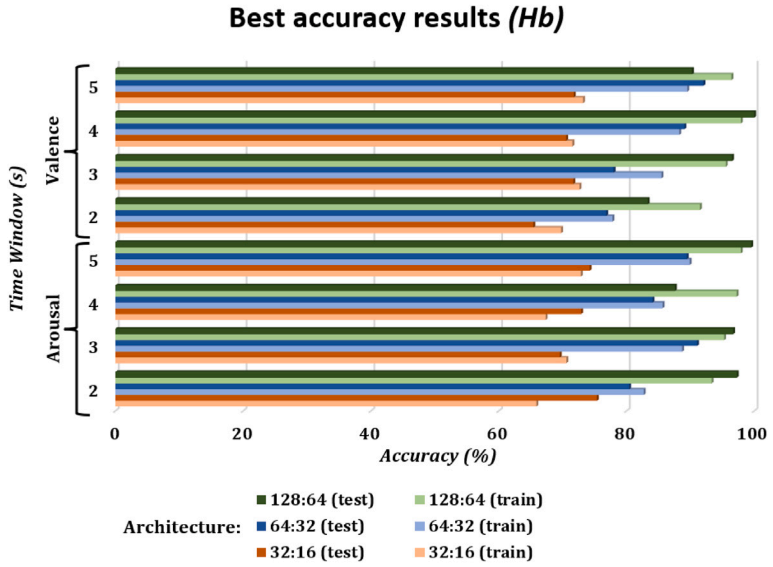


Fig. 9. Graphic representation of the best results obtained for Hb signal.

are worse than the ones obtained with the heaviest architectures, but we expect that the sensor combination improves the final results.

4.1.3. Phase 3: Sensors combination

As detailed before, the two sensors used for the ensemble network are GSR and HbO_2. The frequency features obtained from both sensors are combined in the input layer of the neural network, so the input layer has 16 neurons. The hidden layers have 32 and 16 neurons, respectively, and output layer has 3 neurons (classified classes).

The time windows and batch sizes used for the tests of the ensemble network are based on those parameters with the best accuracy results for each sensor. Those best results that share time window and/or batch size for both sensors are included; moreover, we also analyze the cases of the best results with parameters not shared between both sensors.

So, finally, for *Arousal* tests, the combinations of time window and batch size used are 4–8, 4–16, 5–8 and 5–32. Results are shown in Table 15.

Similarly, in the case of *Valence*, the combinations used are 2–18, 2–16, 3–8, 3–16, 3–32 and 5–32. Results are shown in Table 16.

The results obtained are very interesting: in most cases, the combination of both sensors improves the prediction accuracy (with respect of the results obtained individually), obtaining results exceeding 90% and, in some cases, reaching values higher than 98%. Therefore, the

Table 15
Results obtained with the combined neural network to detect *Arousal* using inputs from HbO_2 and GSR.

HbO_2 + GSR		Train		Test	
Window (s)	BS	Acc (%)	Loss	Acc (%)	Loss
4	8	95.96	0.122	98.60	0.092
4	16	94.82	0.143	89.09	0.6869
5	8	93.82	0.177	99.34	0.099
5	32	91.70	0.217	96.24	0.14

Table 16
Results obtained with the combined neural network to detect *Valence* using inputs from HbO_2 and GSR.

HbO_2 + GSR		Train		Test	
Window (s)	BS	Acc (%)	Loss	Acc (%)	Loss
2	8	85.25	0.369	89.21	0.331
2	16	92.21	0.204	86.00	0.610
3	8	96.25	0.114	99.25	0.065
3	16	94.06	0.171	57.77	4.200
3	32	91.04	0.222	80.41	10.30
5	32	90.00	0.250	83.43	0.715

Table 17
Summary of the Cross-Validation technique applied with 8 different folds.

	Arousal				Valence			
	Repetition 1		Repetition 2		Repetition 1		Repetition 2	
Fold	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss	Acc (%)	Loss
1	98.63	0.220	95.06	0.205	93.01	0.208	96.88	0.119
2	95.45	0.225	95.06	0.298	97.84	0.18	94.63	0.125
3	90.98	0.719	93.20	0.237	98.78	0.104	89.75	0.317
4	100	0.083	87.99	0.645	96.47	0.195	79.14	0.83
5	100	0.103	89.99	0.258	87.87	0.507	89.25	0.333
6	87.99	1.019	97.61	0.343	88.76	0.425	88.94	0.382
7	98.06	0.154	80.40	1.348	79.43	0.77	78.80	0.934
8	89.76	0.856	85.30	1.034	87.35	0.6	83.28	0.678
Mean (%)	95.11	0.4223	90.57	0.546	91.19	0.3736	87.58	0.4647
STD	4.86	0.38	5.79	0.43	6.58	0.2385	6.6799	0.3117
Median (%)	98.06	0.22	93.20	0.32	93.01	0.32	89.40	0.36

Table 18
Neuronal Network final results for *Arousal* classifier.

Class	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	Precision	F1$_{score}$
Low	392	436	0	4	99.5192	98.9899	100	100	99.4924
Medium	140	690	0	2	99.7596	98.5915	100	100	99.2908
High	294	532	6	0	99.2788	100	98.8848	98	98.9899

Table 19
Neuronal Network final results for *Valence* classifier.

Class	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	Precision	F1$_{score}$
Low	666	710	36	2	97.3126	99.7006	95.1743	94.8718	97.2263
Medium	404	988	2	20	98.4441	95.283	99.798	99.5074	97.3494
High	298	1084	8	24	97.7369	92.5466	99.2674	97.3856	94.9045

combination of both sensors with the lighter architecture allows obtaining results similar to those obtained with the more complex architecture; however, the classifier obtained in this section is computationally lighter and facilitates the task of integrating it into an embedded system in future works.

So, we can conclude that, the optimal network for *Arousal* is formed using 5 s time windows, with a *batch size* of 16 and two hidden layers, the first with 32 nodes and the second with 16.

For valuing *Valence* the best network is the one that use 3 s temporal windows, a *batch size* of 16 and has the same structure as the network for classifying *Arousal*, 32 nodes in the first hidden layer and 16 in the second.

4.1.4. Phase 4: Best candidate results

To verify the robustness of the system, the cross-validation technique is applied on the selected classifiers. For this purpose, the dataset has been divided into 8 different, random and non-coincident divisions of test set and training set. Each of these divisions has been named fold. Moreover, these tests have been repeated 2 times, thus obtaining in total 16 different tests for *Arousal* and 16 for *Valence*.

The results obtained are shown in Table 17 for both *Arousal* and *Valence*. To shown the results, the values of the test subset were used. In addition, the mean value between each repetition is extracted, with the standard deviation (STD) and the median.

Certain variations can be observed depending on the fold trained at each moment, but the result demonstrates the robustness of the system. In general, the results show accuracy values above 90% for both Arousal and Valence.

Next, once the robustness and feasibility of the emotion classifier system based on a classical neural network with frequency features extracted from sweat and oxygenated hemoglobin signals has been demonstrated, detailed results on the two systems finally selected are presented.

The results of the various metrics detailed in Section 3 for these two systems, derived from the median results of the first repetition of the cross-validation tests, can be seen in Table 18 for Arousal, and in Table 19 for Valence.

As usual, the accuracy of each class independently is higher than the average accuracy due to the way it is calculated (taking into account, in each case, a two-class system in which the true class is the one evaluated and the false one is a combination of the two remaining ones).

In any case, it can be seen that the results are more than acceptable (above 92% in all cases). As can be seen, the results for Arousal are slightly better than the Valence results; even so, they are acceptable.

The results of the confusion matrices for both classifiers are presented below. Fig. 10 presents the confusion matrix for *Arousal*, and Fig. 11 for *Valence*.

Finally, the ROC curves for both classifiers are shown. Fig. 12 presents the ROC curve for *Arousal*, and Fig. 13 presents the ROC curve for *Valence*.

For the *Arousal* classifier, the area under the ROC curve (AUC) for two of the three classes is 1 (perfect classification), while the value for class 0 (LOW) is 99.7%. And, for the *Valence* classifier, the AUC obtained is 99.8% for class 0 (LOW), 99.4% for class 1 (MID), and 99.5% for class 2 (HIGH).

These results support the conclusions obtained previously: the classifier resulting from the combination of the sweat sensor and the oxygenated hemoglobin sensor obtains more than acceptable results for the detection of *Arousal* and *Valence* values when detecting the emotional state.

4.2. Comparison with previous works

Finally, using the similar work detailed in Section 2, comparative results are presented in terms of classifier type, classes used in the classification, sensors used and results obtained. This comparison can be observed in Table 20.

As can be seen in Table 20, as discussed in Section 2, most of the previous work reviewed includes a two-class classification only. This classification, while acceptable for the detection of certain emotions,

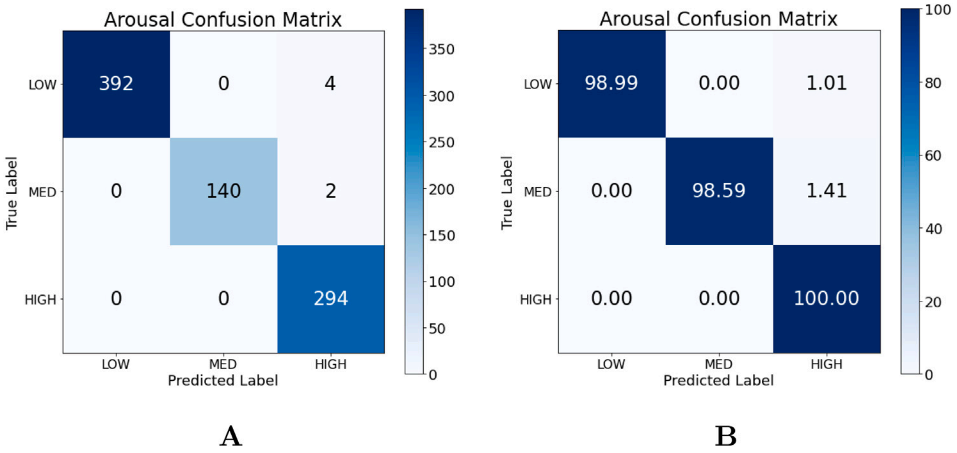


Fig. 10. Confusion matrix obtained for *Arousal*: (A) absolute values, (B) normalized values.

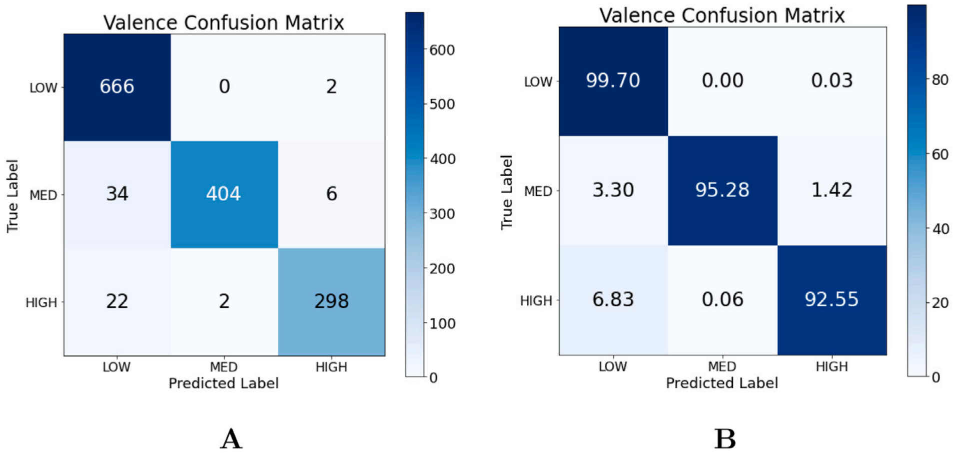


Fig. 11. Confusion matrix obtained for *Valence*: (A) absolute values, (B) normalized values.

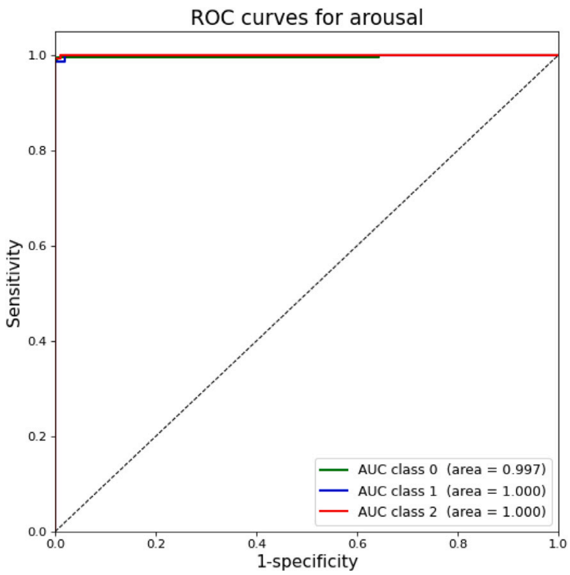


Fig. 12. ROC curve for *Arousal*.

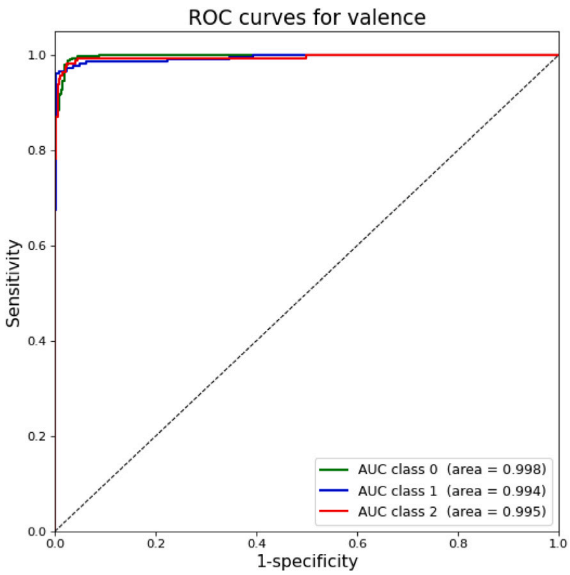


Fig. 13. ROC curve for *Valence*.

Table 20
Works comparison with the results obtained in this work.

Work	Year	Classes	Sensors	Classifier	Results
García et al. (2016)	2016	[3] Low, Medium, High	EEG, EMG and EOG[a]	SVM[d]	Valence: 88.3% Arousal: 90.6%
Liu et al. (2016)	2016	[2]: Low, High	EEG	KNN[e] and RF[f]	Valence: 69.6% Arousal: 71.2%
Li et al. (2016)	2016	[2]: Low, High	EEG	C-RNN[g]	Valence: 72.1% Arousal: 74.1%
Zhang et al. (2016)	2016	[2]: Low, High	EEG	PNN[h]	Valence: 81.2% Arousal: 81.2%
Mirmohamadsadeghi et al. (2016)	2016	[2]: Low, High	ECG and Breathing	SVM[d]	Valence: 74.0% Arousal: 74.0%
Zheng et al. (2017)	2017	[3]: Negative, Neutral, Positive	EEG	KNN[e], LR[i]and SVM[d]	Mean: 79.3%
Girardi et al. (2017)	2017	[2]: Low, High	EEG, GSR and EMG	SVM[d]	Valence: 63.9% Arousal: 58.6%
Lee and Yoo (2018)	2018	[2]: Neutral, Negative	ECG, GSR and SKT[b]	NN[j]	Mean: 92.5%
Lee et al. (2019)	2019	[2]: Low, High	PPG[c]	CNN[k]	Valence: 75.3% Arousal: 76.2%
Sonkusare et al. (2019)	2019	[2]: Low, High	ECG, GSR and SKT[b]	CNN[k]	Mean: 92%
Lee and Yoo (2020)	2020	[2]: Neutral, Negative	ECG, GSR and SKT[b]	B-RNN[l]	Mean: 98.4%
Domínguez-Jiménez et al. (2020)	2020	[3]: Sad, Joy and Neutral	GSR y PPG	SVM	Mean: 100%
Ayata et al. (2020)	2020	[2]: Low, High	SKT, PPG and Breathing	RF, SVM y LR	Valence: 73.08% Arousal: 72.18%
Muñoz-Saavedra et al. (2020)	2020	[3]: Low, Medium, High	ECG and GSR	NN[j]	Valence: 90.4% Arousal: 91.7%
Sepúlveda et al. (2021)	2021	[2]: Low, High	ECG	Ensemble	Valence: 89.1% Arousal: 89.3%
Tabbaa et al. (2021)	2021	[2]: Low, High	ECG, GSR and EOG	SVM	Valence: 90.63% Arousal: 84.38%
Hammad and Monkaresi (2022)	2022	[2]: Low, High	ECG	CNN, DNN	Valence: 76.19% Arousal: 80.95%
Jalal and Peer (2022)	2022	[4]: Angry, Happy, relaxed, sad	GSR, Breathing and Blood pressure	CNN	Mean: 84.20%
This work	–	[3]: Low, Medium and High	PPG and GSR	NN	Valence (average): 84.27% Arousal (average): 91.82% Valence (best): 98.72% Arousal (best): 98.63%

[a] EOG: Electrooculography
[b] C-RNN: Convolutional Recurrent Neural Network
[c] SKT: Skin Temperature
[d] PNN: Probabilistic Neural Network
[e] PPG: Photoplethysmography (Blood Pressure)
[f] LR: Logic Regression
[g] SVM: Supported Vector Machine
[h] NN: Classical MLP Neural Network
[i] KNN: k-nearest neighbors
[j] CNN: Convolutional Neural Network
[k] RF: Random Forest
[l] B-RNN: Bidirectional Recurrent Neural Network

may cause confusion if the range of emotions is extended beyond the classic and most salient emotions.

These are the cases of the works developed by Ayata et al. (2020), Girardi et al. (2017), Lee et al. (2019), Lee and Yoo (2018, 2020), Li et al. (2016), Liu et al. (2016), Mirmohamadsadeghi et al. (2016), Sepúlveda et al. (2021), Sonkusare et al. (2019), Tabbaa et al. (2021) and Zhang et al. (2016).

In our work, a classification is made on three classes of both Arousal and Valence. This, in theory, should cause the prediction results to be lower than those obtained in works with fewer classes; however, this is only the case if we look at the average values of our system (not the best ones) and compare with Lee and Yoo (2018, 2020), Sepúlveda et al. (2021) and Sonkusare et al. (2019). For the other 9 works, despite classifying with only two classes, they obtain worse results than those obtained in this work.

Moreover, if we look closely at those papers with two classes that obtain better accuracy, there are certain extenuating circumstances:

- For the case of Lee and Yoo (2018), the accuracy value provided is the mean (it does not distinguish between *Arousal* and *Valence*), obtaining a mean value of 92.5% (improving the value of 88.05%, obtained as the mean of the final evaluations in our work). However, this work uses three sensors to classify emotions (as opposed to the two finally used in our work). And, likewise, the best result obtained for our work far exceeds that percentage (more than 98%).
- In the case of Sonkusare et al. (2019), with an average accuracy of 92%, something similar to the previous case occurs: it uses three sensors and does not present the results divided.
- For the case of Lee and Yoo (2020), the results obtained rival the best results of this work (98.4% compared to the average of 98.675% of the best case of this work). Again, in addition to the fact that it classifies only two classes, it also uses a third sensor. Moreover, in this case there is another circumstance to be taken into account: the type of classifier used (Bidirectional RNN) is far more complex than the one used in this work. Previous work

by this research group has demonstrated that the computational complexity of an RNN compared to a system with a classical neural network and prior feature extraction is practically 10 times higher when run in an environment without a graphical accelerator (Escobar-Linero, Luna-Perejón, Muñoz-Saavedra, Sevillano, & Domínguez-Morales, 2022). Therefore, as this work uses a more complex classifier than a basic RNN, the computational complexity will be more than 10 times that the one needed to run our classifier.
- Finally, for the case of Sepúlveda et al. (2021), the results obtained are lower than the best in our work (89.1%–89.3% versus 98.72%–98.63%), although slightly better than those obtained in the mean of the final tests of our work. Even so, the main drawback, apart from only classifying two classes, is that the neural network used for classification is an ensemble of several independent networks, which multiplies the complexity of the system used.

Then, if we analyze those works in which three classes are classified, we find Domínguez-Jiménez et al. (2020), García et al. (2016) and Zheng et al. (2017). There is a fourth paper in this case, which is Muñoz-Saavedra et al. (2020), although this paper corresponds to a previous study carried out by the main authors of this paper. In that previous work, the DEAP dataset was also used (like most of the works with which we compare ourselves) with three classes and two sensors, obtaining promising results (90.4% for *Valence* and 91.7% for *Arousal* for the best case). If we compare the work presented here with that previous work, we find several significant differences:

- The dataset used for this work has been collected by this research group using a device and a protocol developed specifically for the occasion. This is because we wanted to demonstrate that the results obtained with a commercial device (DEAP dataset) were applicable to a customized, low-power device so that, after this work, we could integrate the classifier in the embedded device (this could not be done in the case of the commercial device used in the DEAP dataset).
- The development and optimization process of the classifier carried out for this work is much more exhaustive and robust than the one used for the previous work. The use of techniques such as cross-validation and the training of networks combining several sensors are techniques that were not performed in the previous work (in which one sensor was used to classify *Arousal* and another to classify *Valence*).
- The features extracted in this work, unlike the previous work, are based on a previous study performed over previous work and those who gave the best results in the previous work developed by this group.
- The best results obtained in this work improve on those obtained in the previous work (shown in Table 20). It is true that the average results of this work are lower than the best results obtained in previous work; however, the average results of the previous work did not exceed 85%.

Finally, comparing ourselves with the works that classify three classes (except for the previous work of this group), we can make the following analysis:

- For the case of Domínguez-Jiménez et al. (2020), the classification of this work does not follow the two-dimensional emotional theory of *Arousal* and *Valence*, but establishes three clearly differentiated emotions and classifies based on them. If we determine the number of classes of our classifier based on how many different emotions it can distinguish, we can indicate that up to 9 different categories would be classified (3 possible values of *Valence* × 3 possible values of *Arousal*). Therefore, although this work has been indicated as a 3-class classifier, the information it provides is inferior even to a 2-class classifier using two-dimensional emotional theory (since, in that case, up to 4 emotional classifications could be distinguished).
- For the case of Zheng et al. (2017), three different classifiers are used, and none of them is a neural network. Moreover, the results obtained in this work are much lower that the ones obtained in our work (79.3% versus 98.63%–98.72% for the best case).
- And, finally, for the case of García et al. (2016), this is the work with the most similarities with our work. In this work, three classes are used (same as our classifier) and the results obtained are very similar to those obtained in our work in average (although the best case of our work surpasses it a 9%). In this case, the sensors used are not similar to ours: while we use the GSR and PPG signal, this work uses EEG (discarded by us in the previous work due to poor results), EMG and EOG. It is interesting to note that, using a classifier based on SVM, the results are quite good. Therefore, even if our work obtains better results, we will analyze the sensors used in this work for future developments to see if improvements in the classifier are achieved. Finally, it should be noted that the purpose of our work is to integrate the classifier in a non-invasive wearable system; but, in the case of this work, the device should be placed on the head to capture oculography information, facial muscle activity and electrical activity of the cerebral cortex. Therefore, our system is much less invasive than the one indicated in this work.

In summary, an analysis of the most important studies of recent years shows that very few studies actually distinguish between the three classes and, among those that do, the results obtained are inferior to those obtained in our study. Even so, for future work we will carry out more exhaustive comparisons with different types of classifiers and sensors according to what has been observed in previous works.

5. Conclusions

The affective or emotional state of a person can be measured by variations of physiological signals and is related to the two-dimensional emotional theory theorized by Mehrabian and Russell (1974).

In this work, a non-invasive wearable device is developed, composed of several physiological sensors, which is capable of capturing such information in real time and storing it in a computer.

With this, a protocol is designed, endorsed by an ethical committee, in which emotions are induced in a group of participants using audiovisual information while the physiological data information registered in real time by the wearable device is recorded.

This collected dataset is labeled based on *Arousal* and *Valence* values according to the two-dimensional emotional theory, and an automatic classifier is designed based on neural networks with prior extraction of frequency characteristics of the physiological signals.

For the design of this classifier, a meticulous process consisting of four phases is followed: first, all the sensors are evaluated; second, the most suitable sensors are determined; third, combinations of different sensors are performed; and, finally, the robustness of the final classifier is evaluated by applying cross-validation techniques on the developed neural network.

The test subset classification results of the designed neural network exceed, for the best case, 98% for both *Arousal* and *Valence*. If we compare the results obtained with previous works, improvements are observed in the number of classes used by the classifier, its accuracy, system architecture complexity, and the non-invasive wearable.

These results demonstrate the feasibility of using a wearable system for the classification of the user's emotional state. Furthermore, the results obtained, the network simplicity, and the non-invasive wearable allow our system to be analyzed in order to integrate it into the embedded system in future work.

CRediT authorship contribution statement

Luis Muñoz-Saavedra: Software tests, Writing. **Elena Escobar-Linero:** Tests, Writing. **Lourdes Miró-Amarante:** Conceptualization, Methodology, Writing. **M. Rocío Bohórquez:** Protocol, Data collection, Interviews. **Manuel Domínguez-Morales:** Conceptualization, Methodology, Writing, Coordination, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Spanish AEI project MINDROB, Spain (PID2019-105556GB-C33), and by the Andalusian Regional I+D+i FEDER, Spain Projects DAFNE US-1381619 and MSF-PHIA, Spain US-1263715.

References

Ardern, C. L., Kvist, J., & Webster, K. E. (2016). Psychological aspects of anterior cruciate ligament injuries. *Operative Techniques in Sports Medicine*, *24*, 77–83.

Ayata, D., Yaslan, Y., & Kamasak, M. E. (2020). Emotion recognition from multimodal physiological signals for emotion aware healthcare systems. *Journal of Medical and Biological Engineering*, *40*, 149–157.

Biagetti, G., Crippa, P., Falaschetti, L., Tanoni, G., & Turchetti, C. (2018). A comparative study of machine learning algorithms for physiological signal classification. *Procedia Computer Science*, *126*, 1977–1984.

Civit-Masot, J., Domínguez-Morales, M. J., Vicente-Díaz, S., & Civit, A. (2020). Dual machine-learning system to aid glaucoma diagnosis using disc and cup feature extraction. *IEEE Access*, *8*, 127519–127529.

Civit-Masot, J., Luna-Perejón, F., Corral, J. M. R., Domínguez-Morales, M., Morgado-Estevez, A., & Civit, A. (2021). A study on the use of edge TPUs for eye fundus image segmentation. *Engineering Applications of Artificial Intelligence*, *104*, Article 104384.

Civit-Masot, J., Luna-Perejón, F., Domínguez Morales, M., & Civit, A. (2020). Deep learning system for COVID-19 diagnosis aid using X-ray pulmonary images. *Applied Sciences*, *10*, 4640.

Diener, E., & Chan, M. Y. (2011). Happy people live longer: Subjective well-being contributes to health and longevity. *Applied Psychology: Health and Well-Being*, *3*, 1–43.

DiMaria, C. H., Peroni, C., & Sarracino, F. (2020). Happiness matters: Productivity gains from subjective well-being. *Journal of Happiness Studies*, *21*, 139–160.

Domínguez-Jiménez, J. A., Campo-Landines, K. C., Martínez-Santos, J. C., Delahoz, E. J., & Contreras-Ortiz, S. H. (2020). A machine learning model for emotion recognition from physiological signals. *Biomedical Signal Processing and Control*, *55*, Article 101646.

Domínguez-Morales, M. J., Luna-Perejón, F., Miró-Amarante, L., Hernández-Velázquez, M., & Sevillano-Ramos, J. L. (2019). Smart footwear insole for recognition of foot pronation and supination using neural networks. *Applied Sciences*, *9*, 3970.

Escobar-Linero, E., Domínguez-Morales, M., & Sevillano, J. L. (2022). Worker's physical fatigue classification using neural networks. *Expert Systems with Applications*, *198*, Article 116784.

Escobar-Linero, E., Luna-Perejón, F., Muñoz-Saavedra, L., Sevillano, J. L., & Domínguez-Morales, M. (2022). On the feature extraction process in machine learning. An experimental study about guided versus non-guided process in falling detection systems. *Engineering Applications of Artificial Intelligence*, *114*, Article 105170.

Frey, B. S., & Stutzer, A. (2016). Policy consequences of happiness research. In *Policies for Happiness* (pp. 21–35). Oxford: Oxford University Press.

García, H. F., Álvarez, M. A., & Orozco, Á. A. (2016). Gaussian process dynamical models for multimodal affect recognition. In *2016 38th A nnual international conference of the ieee engineering in medicine and biology society* (pp. 850–853). IEEE.

Girardi, D., Lanubile, F., & Novielli, N. (2017). Emotion detection using noninvasive low cost sensors. In *2017 seventh international conference on affective computing and intelligent interaction* (pp. 125–130). IEEE.

Hammad, D. S., & Monkaresi, H. (2022). ECG-based emotion detection via parallel-extraction of temporal and spatial features using convolutional neural network. *Traitement du Signal*, *39*.

Harmon-Jones, C., Bastian, B., & Harmon-Jones, E. (2016). The discrete emotions questionnaire: A new tool for measuring state self-reported emotions. *PLoS One*, *11*, Article e0159915.

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve? *Emergency Medicine Journal*, *34*, 357–359.

Howell, R. T., Kern, M. L., & Lyubomirsky, S. (2007). Health benefits: Meta-analytically determining the impact of well-being on objective health outcomes. *Health Psychology Review*, *1*, 83–136.

Hu, T. -Y., Xie, X., & Li, J. (2013). Negative or positive? The effect of emotion and mood on risky driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *16*, 29–40.

Jalal, L., & Peer, A. (2022). Emotion recognition from physiological signals using continuous wavelet transform and deep learning. In *International conference on human-computer interaction* (pp. 88–99). Springer.

Kwon, Y. -H., Shin, S. -B., & Kim, S. -D. (2018). Electroencephalography based fusion two-dimensional (2D)-convolution neural networks (CNN) model for emotion recognition system. *Sensors*, *18*, 1383.

Lee, M. S., Lee, Y. K., Pae, D. S., Lim, M. T., Kim, D. W., & Kang, T. K. (2019). Fast emotion recognition based on single pulse PPG signal with convolutional neural network. *Applied Sciences*, *9*, 3355.

Lee, J., & Yoo, S. K. (2018). Design of user-customized negative emotion classifier based on feature selection using physiological signal sensors. *Sensors*, *18*, 4253.

Lee, J., & Yoo, S. K. (2020). Recognition of negative emotion using long short-term memory with bio-signal feature compression. *Sensors*, *20*, 573.

Li, X., Song, D., Zhang, P., Yu, G., Hou, Y., & Hu, B. (2016). Emotion recognition from multi-channel EEG data through convolutional recurrent neural network. In *2016 IEEE international conference on bioinformatics and biomedicine* (pp. 352–359). IEEE.

Liu, J., Meng, H., Nandi, A., & Li, M. (2016). Emotion detection from EEG recordings. In *2016 12th international conference on natural computation, fuzzy systems and knowledge discovery* (pp. 1722–1727). IEEE.

Luna-Perejón, F., Domínguez-Morales, M. J., & Civit-Balcells, A. (2019). Wearable fall detector using recurrent neural networks. *Sensors*, *19*, 4885.

Luna-Perejón, F., Domínguez-Morales, M., Gutiérrez-Galán, D., & Civit-Balcells, A. (2020). Low-power embedded system for gait classification using neural networks. *Journal of Low Power Electronics and Applications*, *10*, 14.

Luna-Perejón, F., Muñoz-Saavedra, L., Civit-Masot, J., Civit, A., & Domínguez-Morales, M. (2021). AnkFall—Falls, falling risks and daily-life activities dataset with an ankle-placed accelerometer and training using recurrent neural networks. *Sensors*, *21*, 1889.

Megías, C. F., Mateos, J. C. P., Ribaudi, J. S., & Fernández-Abascal, E. G. (2011). Validación española de una batería de películas para inducir emociones. *Psicothema*, 778–785.

Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology*. The MIT Press.

Mirmohamadsadeghi, L., Yazdani, A., & Vesin, J. -M. (2016). Using cardio-respiratory signals to recognize emotions elicited by watching music video clips. In *2016 IEEE 18th international workshop on multimedia signal processing* (pp. 1–5). IEEE.

Muñoz-Saavedra, L., Luna-Perejón, F., Civit-Masot, J., Miró-Amarante, L., Civit, A., & Domínguez-Morales, M. (2020). Affective state assistant for helping users with cognition disabilities using neural networks. *Electronics*, *9*, 1843.

Neumann, W. P., & Dul, J. (2010). Human factors: Spanning the gap between OM and HRM. *International Journal of Operations & Production Management*.

Ødegaard, F., & Roos, P. (2014). Measuring the contribution of workers' health and psychosocial work-environment on production efficiency. *Production and Operations Management*, *23*, 2191–2208.

Santamaria-Granados, L., Munoz-Organero, M., Ramirez-Gonzalez, G., Abdulhay, E., & Arunkumar, N. (2018). Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS). *IEEE Access*, *7*, 57–67.

Sepúlveda, A., Castillo, F., Palma, C., & Rodriguez-Fernandez, M. (2021). Emotion recognition from ECG signals using wavelet scattering and machine learning. *Applied Sciences*, *11*.

Sokolova, M., et al. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*, 427–437.

Sonkusare, S., Ahmedt-Aristizabal, D., Aburn, M. J., Nguyen, V. T., Pang, T., Frydman, S., et al. (2019). Detecting changes in facial temperature induced by a sudden auditory stimulus based on deep learning-assisted face tracking. *Scientific Reports*, *9*, 1–11.

Tabbaa, L., Searle, R., Bafti, S. M., Hossain, M. M., Intarasisrisawat, J., Glancy, M., et al. (2021). VREED: Virtual reality emotion recognition dataset using eye tracking & physiological measures. In *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*: *vol. 5*, (pp. 1–20). NY, USA: ACM New York.

Te Wierike, S., van der Sluis, v. d. A., van den Akker-Scheek, I., Elferink-Gemser, M., & Visscher, C. (2013). Psychosocial factors influencing the recovery of athletes with anterior cruciate ligament injury: A systematic review. *Scandinavian Journal of Medicine & Science in Sports*, *23*, 527–540.

Tenney, E. R., Poole, J. M., & Diener, E. (2016). Does positivity enhance work performance?: Why, when, and what we don't know. *Research in Organizational Behavior*, *36*, 27–46.

Ton, Z. (2014). *The good jobs strategy: How the smartest companies invest in employees to lower costs and boost profits*. Houghton Mifflin Harcourt.

Wiese-Bjornstal, D. M., Smith, A. M., Shaffer, S. M., & Morrey, M. A. (1998). An integrated model of response to sport injury: Psychological and sociological dynamics. *Journal of Applied Sport Psychology*, *10*, 46–69.

Zhang, J., Chen, M., Hu, S., Cao, Y., & Kozma, R. (2016). PNN for EEG-based emotion recognition. In *2016 IEEE international conference on systems, man, and cybernetics* (pp. 002319–002323). IEEE.

Zheng, W. -L., Zhu, J. -Y., & Lu, B. -L. (2017). Identifying stable patterns over time for emotion recognition from EEG. *IEEE Transactions on Affective Computing*.