Expert Systems With Applications 198 (2022) 116784

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Worker's physical fatigue classification using neural networks

Elena Escobar-Linero [a], Manuel Domínguez-Morales [b,*], José Luis Sevillano [c]

[a] Biomedical Engineering, Architecture and Computer Technology Department (ATC), Robotics and Technology of Computers Lab (RTC), E.T.S. Ingeniería Informática, Avda. Reina Mercedes s/n, Universidad de Sevilla, Seville, 41012, Spain

[b] 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, Avda. Reina Mercedes s/n, Universidad de Sevilla, Seville, 41012, Spain

[c] Physics, Architecture and Computer Technology Department (ATC), Robotics and Technology of Computers Lab (RTC), Computer Engineering Research Institute (I3US), E.T.S. Ingeniería Informática, Avda. Reina Mercedes s/n, Universidad de Sevilla, Seville, 41012, Spain

ARTICLE INFO

Keywords:
Deep Learning
Fatigue
Physical activity

ABSTRACT

Physical fatigue is not only an indication of the user's physical condition and/or need for sleep or rest, but can also be a significant symptom of various diseases. This fatigue affects the performance of workers in jobs that involve some continuous physical activity, and is the cause of a large proportion of accidents at work. The physical fatigue is commonly measured by the perceived exertion (RPE). Many previous studies have attempted to continuously monitor workers in order to detect the level of fatigue and prevent these accidents, but most have used invasive sensors that are difficult to place and prevent the worker from performing their tasks correctly. Other works use activity measurement sensors such as accelerometers, but the large amount of information obtained is difficult to analyse in order to extract the characteristics of each fatigue state. In this work, we use a dataset that contains data from inertial sensors of several workers performing various activities during their working day, labelled every 10 min based on their level of fatigue using questionnaires and the Borg fatigue scale. Applying Machine Learning techniques, we design, develop and test a system based on a neural network capable of classifying the variation of fatigue caused by the physical activity collected every 10 min; for this purpose, a feature extraction is performed after the time decomposition done with the Discrete Wavelet Transform (DWT). The results show that the proposed system has an accuracy higher than 92% for all the cases, being viable for its application in the proposed scenario.

1. Introduction

Fatigue is a complex, multidimensional term, without a standard and universal definition, since it can be attributed to many factors (Aghdam et al., 2019). According to Health Safety Executive, fatigue can be defined as "the result of prolonged mental or physic exertion, which can affect people's performance and impair their mental alertness" (Health and Safety Executive, 2006). Another definition of fatigue is "a state of feeling tired, weary, or sleepy that results from prolonged mental and physical work, extended periods of anxiety, exposure to harsh environment, or loss of sleep" (Sadeghniiat & Yazdi, 2015) or "any exercise-induced reduction in the maximal capacity to generate force or power output" (Vøllestad, 1997). As noted, fatigue results from two main types: physical and mental fatigue. Physical fatigue consists of a reduction in capacity to perform physical work as a function of preceding physical exertion, which influences performance (Gawron, French, & Funke, 2001).

Fatigue is also related to a number of diseases such as multiple sclerosis (MS) (Goldenberg, 2012), chronic fatigue syndrome (CFS) (Greenberg & Frid, 2006), fibromyalgia (Busch et al., 2011), myasthenia gravis (O'Connor et al., 2020), chronic obstructive pulmonary disease (COPD) (Al-shair et al., 2016) and Eaton–Lambert syndrome (Lang et al., 1981). The pathology, causes or consequences of these diseases-related fatigue vary depending on multiple factors. However, in this paper we will focus on work-related fatigue, that is, fatigue suffered in a work context. Reduction of work-related fatigue is an important issue as it can lead to a reduction of injuries or accidents in the workplace, less work absenteeism and an improvement of work performance (Masala et al., 2017).

In manufacturing industry, physical fatigue is a common issue, mostly because of high demand tasks, long duty periods and accumulative sleep (Sadeghniiat & Yazdi, 2015). Principal causes that influence the onset of fatigue in the workplace are individual factors, such as

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

E-mail addresses: eescobar@us.es (E. Escobar-Linero), mjdominguez@us.es (M. Domínguez-Morales), jlsevillano@us.es (J.L. Sevillano).

https://doi.org/10.1016/j.eswa.2022.116784
Received 13 August 2021; Received in revised form 22 December 2021; Accepted 26 February 2022
Available online 18 March 2022

socioeconomic, psychological and work load. The latter can be assessed in three categories, including physical load, mental load and environmental load (Lambay et al., 2021; Yung et al., 2014). Work-related physical fatigue can lead to short-term problems, such as discomfort, decreasing level of performance, strength, productivity and quality of work, as well as increasing errors and work accidents, and it can also lead to long-term issues, like cardiovascular diseases or CFS (Aghdam et al., 2019; Lambay et al., 2021).

Despite technological and industry improvements, work-related fatigue still prevails. On the report of National Safety Council (NSC), in 2018 two-third parts of United States labor force suffered from work-related fatigue, i.e., approximately 107 million out of the 160 million workers were affected. A research in 2007 stipulated that workers with fatigue, associated with lost of productivity and other issues, cost to the U.S. 101 billion dollars annually (Ricci et al., 2007). Moreover, it is demonstrated that 13% of workplace injuries can be attributed to fatigue (National Safety Council, 2020). In Spain, 30.8% of workers suffer from fatigue, and in Europe, 3.2% of people of 15–64 years that worked during the past year had one or more accidents at work in the past 12 months, where approximately 70% of the non-mortal accidents resulted from loss of control or a fall due to stress or work-related fatigue (European Commission, 2010).

Therefore, for all the previous reasons and in order to avoid them, it is important to detect and measure work-related fatigue, so that injuries, accidents or diseases can be prevented and recommendations to reduce or avoid exertion can be done.

In order to detect and measure fatigue, there are objective methods, based on the analysis of body parts that exert the force required during the task, and subjective methods, based on analysis of fatigue through rating scales and questionnaires made to individuals for assessing perceived exertion (Sedighi Maman et al., 2017). Perceived exertion or perceived effort is the most common measure to quantify fatigue, and it is defined by Borg as "the perception of how the body is working during an exercise, a 'Gestalt' based on many sensory cues and perceptions" (Borg, 2007). In work-related fatigue studies, the assessment of symptoms and discomforts should be done subjectively and, for this reason, three different techniques of psychophysical measurement of perceived exertion have been developed in the last decades: ratio scaling, category scaling and acceptability scaling (Borg, 1990). Ratio scaling have been developed to obtain methods with same metric qualities as those used in physics and physiology, i.e., methods with an absolute zero and same distance between scale values (Gamberale, 1985). Magnitude Estimation is one of the most famous scales, created by Stevens in 1975, which is a perceptual scaling technique where participants are instructed to assign numbers in relation to perceptual intensities (Borg, 2007). Years later, Borg developed the Rating of Perceived Exertion scale (RPE), which measures subject's effort and exertion while performing a physical task, through a ranked score from 6 (no exertion at all) to 20 (maximal exertion) (Borg, 1982). The same author, years afterwards, developed a scale of categories to assess perceived exertion, CR10 scale, ranked from 0 to 10 with verbal anchors, where 10 indicates a extremely strong perceived exertion and categorizes as 'maximal' (Watt & Grove, 1993). RPE scale is considered best for simple applied studies of perceived exertion and for physical intensities prediction, while CR10 scale is more suitable to assess subjective symptoms (Borg, 1982). Other category scales are CR100 (centiMax), ranked from 0 to 100, or OMNI-RPE scale, with a ranked score 0–10 adding mode-specific pictures (Ritchie, 2012).

As mentioned before, another tool to assess perceived exertion is through questionnaires. Some examples of them are Fatigue Severity Scale (FSS), which is a 9-item self-report questionnaire developed for diseases-related fatigue control (Krupp et al., 1989), Multidimensional Fatigue Inventory (MFI-20), a 20-item questionnaire with 5 sub-scales: general fatigue, physical fatigue, reduced activity, reduced motivation and mental fatigue (Bunevicius et al., 2011), and Chadler Fatigue Scale (CFQ), which measures physical and psychological fatigue, with an 11-item questionnaire rating as 0 (better than usual), 1 (not worse than usual), 2 (worse than usual), 3 (much worse than usual) (Aghdam et al., 2019).

On the other hand, in the past few years, there has been an increase in the use of wearable devices, capable to provide monitoring in real time, recording, and communication of individuals' physical and environmental exposures. These devices have been developed in shapes of watches, wrist bands, glasses, jewellery, skin patches and even smart textiles (Seneviratne, Hu, Nguyen, Lan, Khalifa, Thilakarathna, Hassan, & Seneviratne, 2017). One of the most important element in wearable devices are sensors and their use is widespread mostly across sports environment, and with improvement in semiconductor technology, monitoring a full range of parameters is a possibility and the application of wearable devices in medicine is closer (Mostafa et al., 2017). Some wearable devices to monitor health parameters and biosignals are ECG monitors, blood pressure monitors and biosensors, among others (Lee & Lee, 2020). Although the availability and use of these devices in health field is slow-going because of the need of validation in the context of different pathologies, multiple works have demonstrated the feasibility of its use in research projects to predict, monitor or assess several diseases and disabilities (Beniczky et al., 2021; Dominguez-Morales et al., 2019).

In addition to ratings of perceived exertion to assess work-related fatigue as mentioned earlier, the use of wearable devices to monitor body parameters is under research and appears to be a possibility, so that physical exposures of subjects in the workplace can be quantified, monitoring brain activation using electroencephalography (EEG) or monitoring changes in local muscle with electromyography (EMG) (Dong et al., 2014). However, EEG- and EMG-based methods to measure physical fatigue are intrusive, because both of them require several electrodes and are unlikely to resist operational environment as they are most suitable for stationary tasks, besides being expensive technologies (Baghdadi et al., 2019; Balkin et al., 2011). For this reason, researchers are focusing on the possibility of using non-intrusive wearable sensors to workplaces in order to monitor physical activity and mobility, since these devices are cheaper and easier to use (Fu et al., 2019). One of the most used wearable sensors to detect subjects movement is the inertial measurement unit (IMU) to obtain angular velocity and acceleration data.

In the last years, many work-related investigations use IMUs for ergonomic evaluation (Vignais et al., 2013), posture analysis (Battini et al., 2014), musculoskeletal disorders evaluation (Tee et al., 2017), assessment of body motion and lifting risks while doing handling tasks (Barim et al., 2019) or falling risk during daily activities (Luna-Perejón et al., 2019, 2021), among others. Moreover, recently, more studies using IMUs have addressed to the detection of physical activity and fatigue (Karvekar et al., 2021; Lamooki et al., 2020; Schmidt et al., 2016).

In order to develop generalist systems capable of detecting the fatigue state by processing large amounts of data, the use of Artificial Intelligence has been extended. Many previous works have made use of Machine Learning (ML) or Deep Learning (DL) techniques applied on physiological signals to develop systems capable of more easily extracting relevant features from the dataset (Al-Saegh et al., 2021; Lih et al., 2020; Muñoz-Saavedra et al., 2020).

Regarding DL applied to fatigue classification, in 2017 Sedighi Maman et al. used IMUs to obtain acceleration and jerk data from participants performing manufacturing tasks, together with Heart Rate (HR) data and RPE values to implement a *Least Absolute Shrinkage and Selection Operator* (LASSO) model to select features from data and apply regression and logistic models to estimate physical fatigue level. Years later, the previous author proposed a framework centred on detection, identification, diagnosis and recovery from fatigue, in order to quantify and predict changes in workers' performance (Sedighi Maman et al., 2020). On the other hand, Karvekar et al. (2019) used

accelerometers embedded on smartphones in order to measure motion and gait parameters of the participants, along with RPE values collected to label the data, and they applied a ML algorithm (SVM model) to classify subjects' fatigue level, as well as (Zhang et al., 2013), Baghdadi et al. (2018), and Kuschan and Krüger (2021) who also used a SVM model to detect physical fatigue. Another research focused on this area have measured cycle acceleration of workers doing several tasks with IMUs for fatigue detection, applying Statistical Process Control (SPC) techniques (Lamooki et al., 2020). More recently, Lambay et al. (2021) used the dataset from (Sedighi Maman et al., 2017) in order to perform fatigue prediction for manual material handling task. All of these studies facilitate the implementation of a proactive approach in continuous monitoring of operators' fatigue level, which may increase work performance and reduce risks mentioned earlier (Darbandy et al., 2020; Nasirzadeh et al., 2020).

The aim of this work is to study the relationship between physical activity and fatigue, and to develop a classifier system capable of estimating the variation of physical fatigue in periods of 10 min. In this way, every 10 min of work, it would be possible to know whether the worker is more (or less) tired and to what extent. In future works, integrating the classifier in an embedded systems, personalized information could be provided in real time to the worker so that he/she knows his/her degree of fatigue and, in this way, he/she can take breaks when the fatigue reaches dangerous levels.

On the other hand, for future work, a similar method could be used with people with pathologies that cause a sudden increase in physical fatigue in certain circumstances (such as chronic fatigue syndrome) although of course the characteristics of this kind of fatigue are different to worker's physical fatigue. Anyway, with this approach, these people could be helped to plan the tasks to be carried out based on the degree of accumulated fatigue detected by the system.

As noted, multiple techniques can be applied to predict work-related fatigue state. In this study, we have used a time-series dataset which contains acceleration and jerk data from four IMUs disposed on seven participants and HR data, while they were doing three different tasks. With this dataset, we have applied a Discrete Wavelet Transform (DWT) and extracted adequate features, which are the inputs of a Sequential Neural Network, and with RPE window-variation values as target of the network, the perceived exertion level has been predicted. In Section 2 of this study, we review the physiological signals measured, along with the manufacturing tasks performed from the dataset obtained, as well as we detail data pre-processing, feature extraction and the neural network implemented. In the last point of this section, we explain the evaluation mechanism followed. Section 3 includes our results and a discussion of work-related fatigue prediction; and, finally, in Section 4 the conclusions of the work are presented.

2. Materials and methods

In this section, the proposed system for classifying the fatigue state of the user is described. This section focuses on the description of the techniques and decisions applied in the development and testing of the classification system, while the results of their application will be presented in Section 3.

So, first, the dataset used for this work is presented in the first subsection; after that, the internal structure of the classification system is explained in the second subsection; in the third place, the mechanisms and metrics used to evaluate the proposed system are described; and, finally, the comparative evaluation of results with previous work and a brief state of the art about them is detailed. The graphical abstract that better describes the complete work is shown in Fig. 1.

2.1. Dataset

Inertial motion of body parts is a common measure to assess physical fatigue, especially in an industrial environment, since, as mentioned earlier, physiological measurements based on EMG or EEG activity are often impractical to implement in work conditions: they can lose contact over time and can only measure the activity of particular body parts (Jiang et al., 2021). To obtain this kinematic data, IMUs are the most commonly used. These are devices that include built-in sensors: accelerometers, gyroscopes and magnetometers along three axes (Ahamed et al., 2021). With this wearable device, there are many advantages such as low cost, good portability, small size, non-intrusiveness, versatility and ease of operation (Dzeng et al., 2014).

In fact, a wide variety of studies demonstrate the feasibility of using accelerometers to detect physical fatigue, although most of them are addressed to fall detection, to detect binary fatigue (fatigue vs. non-fatigue), gait movement or running analysis (Atiya et al., 2021; Buckley et al., 2017; Strohrmann et al., 2012; Zhang et al., 2013). Movement parameters obtained with an inertial measurement unit represents acceleration, angular velocity and magnetic orientation over time. Physical fatigue has been linked to a reduction in motor control and in strength capacity, which is assessed by measuring jerk. Jerk is obtained from the first time-derivative of acceleration, and several studies, e.g., Zhang et al. (2019), demonstrated the feasibility of using jerk metric to detect physical fatigue. In order to capture fatigue from physical tasks, the best range of movement parameters should be measured from the hip, wrist, ankle, and torso (Sedighi Maman et al., 2017; Zhang et al., 2020). Thus, it is known that the use of IMUs have a high potential to be used as a fatigue assessment tool.

The dataset used to train and test the proposed system was obtained from Sedighi Maman et al. (2017), when they investigate the feasibility of using wearable sensors to monitor and detect physical fatigue in working occupations. A brief summary about the sensors used and their body location to monitor physical tasks, an explanation of the tasks performed by the workers and the measures obtained are shown in the next subsections.

2.1.1. Sensors

For the measurement of inertial motion, eight participants, 3 female and 5 female, aged from 18 to 62 years, were instrumented with four IMUs, localized on their ankle, wrist, torso and hip, while they realized three different types of manufacturing tasks. IMU model used for the dataset was the Shimmer3, which contains a low-noise analog accelerometer, a wide range accelerometer and magnetometer, and a digital gyroscope. The data was recorded at a sampling rate of 51.2 Hz, even though after recording, data was cleaned and down-sampled at 25 Hz in order to make it consistent.

2.1.2. Tasks

All eight participants had to complete three physically fatiguing tasks within three one-hour periods each. The first task consisted on part assembly operation (PA), where the subjects had to build assemblies based on visual instructions while in a stationary standing position, so after three hours, physical fatigue is very likely to occur. The second one was a supply pickup and insertion task (SPI), consisting of walking with supplies and bending to a bolt box. Lastly, the third and most physical fatiguing task, named manual material handling (MMH), simulated warehousing operations.

2.1.3. Measures

Acceleration was measured from all four IMUs of each participant while they were performing the previously explained tasks. From this measure, which was recorded every 40 ms, jerk was also calculated as the derivative of acceleration with respect to time. Besides these two measures, participants had to provide in 10 min time-window their subjective exertion value using Borg's RPE scale.

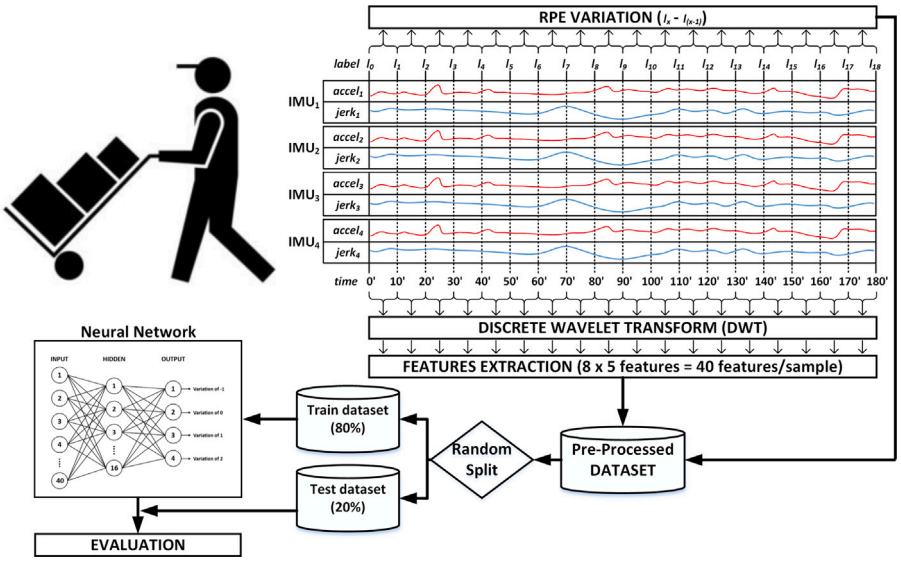

Fig. 1. Graphical abstract of the complete system.

2.2. Perceived exertion classifier

The proposed system for perceived exertion's prediction is based on a Multi-Layer Perceptron Neural Network trained with the data from the previously described dataset. However, in order to train and test the classification system, two previous stages need to be performed: a discrete wavelet transform and a feature extraction step (see Fig. 1).

These two stages, together with the description of the neural network used to classify the perceived exertion or effort, are described next.

2.2.1. Discrete Wavelet Transform

The pre-processing step is very important in order to improve neural network's accuracy. For the dataset employed, data have been compressed using Discrete Wavelet Transform (DWT), so that acceleration and jerk data, measured every 40 ms during 3 h, could be divided into 10-minutes time windows. This time division is based on the tagging process performed for the dataset, which includes a subjective fatigue level questionnaire after every 10 min; so we divide the study on those time windows in order to have valid labels to train and test the classification system.

The selection of frequency approach prior to the feature extraction process in order to assess physical fatigue is due to the good results obtained in fatigue studies, where it is demonstrated that frequency domain features can replace time-domain method (Braccesi et al., 2015). The traditional methods used for features extraction (next stage), as mean or median of the frequency spectrum, are commonly used with Discrete Fourier Transform (DFT). Using this method, the frequency domain is well quantified, but it does not give any timing information; i.e., DFT estimates how much of each frequency exists in the spectrum, but cannot determine when a particular frequency component takes place in time (Chowdhury et al., 2013). This is a good method for periodic analog signal, but for physiological signals like acceleration, where the components vary their periodicity, using DFT is not efficient. For such cases, Discrete Wavelet Transform (DWT) provides simultaneous spectral and temporal information from the signal (Muñoz-Saavedra et al., 2020).

The DWT uses filter banks, which contains high and low frequency filters for the analysis and reconstruction of a signal. In this way, DWT decomposes the signal in approximation coefficients (high frequencies) and detail coefficients (low frequencies), filtering them through scaling and wavelet filters, respectively, and then, the approximation coefficients are subsampled into new coefficients (Mitchell et al., 2013). Moreover, DWT needs a mother wavelet, where, one of the most used is Daubechies wavelet (Sekine et al., 2000). In this investigation, a third-order Daubechies wavelet (db3) is applied, in order to obtain a set of approximation coefficient vectors, A_i, and detail coefficient vectors $D_1, D_2, \ldots, D_i$, at the i_{th} level. So, DWT is more suitable for the system we propose.

As for the labelled information regarding the perceived exertion, which contains scoring values from 6 to 20 (according to Borg scale), it has been converted to the variation values between each RPE time window. This means that, instead of maintain the original labels (which refer to the absolute value of perceived exertion), the new labels contain the RPE variation between the beginning and the end of that 10-minutes time windows (this is the increase in perceived exertion during that time window).

For example: at the beginning of the 10-minutes window the subject has indicated a level of 6 in perceived exertion. During the next 10 min of doing physical tasks, the level of perceived exertion increments to 8 at the end of the time window. So, the variation label used in our system will be 2 for that time window. Hence, perceived effort state is discretized in variation levels of −1, 0, 1, and 2; that is because, for all the users and tasks included in the dataset, the variation RPE from the beginning to the end of a time window does not vary more than 2.

Previous works have used this dataset to train a classification system using the absolute values, but the results have not been very satisfactory (Lambay et al., 2021). This is because the perceived exertion/effort is cumulative and, therefore, if we want to train the system with absolute values, we will have to include the physical activity from the beginning to the current point (not just the 10-minutes time window).

Therefore, the labelling used for this work is more consistent with reality since, by extracting information from a 10-minutes time window, we will only be able to detect the variation in the perceived exertion during those 10 min. However, in order to obtain an absolute value of the user's state in our system, we will have to accumulate the variations in the perceived exertion during all the 10-minutes time windows from the beginning to the current point.

After this step, and before the feature extraction process (described next), the Wavelet Coefficients' values of each time window are normalized. Normalization consists in a re-scale process in order to obtain an output between 0 and 1, known as min–max normalization; where, for each value, the minimal range value is subtracted, and then divided by the feature's range.

2.2.2. Features extraction

The feature extraction is a process of dimension reduction without losing important information, where a large amount of data is reduced in a feature. In this study, in order to train the system and improve neural network's accuracy, frequency features are previously extracted using DWT (described before).

After obtaining wavelet coefficients, features can be extracted. For acceleration signals, several studies have used DWT to calculate the total energy of the wavelet transform decomposition (Ayrulu-Erdem & Barshan, 2011). Specifically, the energy ratio of the DWT coefficients to the total energy can give discriminatory results. E_T represents the total energy at level i, while EDR_A represents the energy ratio of the approximation coefficients vector at level i and EDR_{D_j} refers to the energy ratio of the detail coefficient vector j, where $j = 1, \dots, i$. All of these are given by (Ayrulu-Erdem & Barshan, 2011):

$$E_T = A_i A_i^T + \sum_{j=1}^{i} D_j D_j^T \tag{1}$$

$$EDR_A = \frac{A_i A_i^T}{E_T} \tag{2}$$

$$EDR_{D_j} = \frac{D_i D_j^T}{E_T} \tag{3}$$

Besides these features related to the DWT energy distribution, Ayrulu-Erdem and Barshan (2011) and Mitchell et al. (2013), demonstrated that the combination of the latter and the normalized variances provide the most informative features. Therefore, in this work, the number of features extracted consist of the normalized variances of the acceleration and jerk data, which gives a number of 2 features, and the EDR_A and EDR_{D_j} features, which are equal to $i + 1$ number of features at i_{th} level. Combining both variances and energy features of acceleration and jerk data gives a total of $2 \times (i + 2)$ features at level i. Finally, taking in account the four IMUs located in subjects' body, the total number of features for a participant is $8 \times (i+2)$. In this work, level 3 of the decomposition is used, so the total amount of coefficients per sample are $8 \times (3 + 2) = 40$.

2.2.3. Neural network model

For the prediction of perceived exertion state, we use a feed-forward Multilayer Perceptron (MLP) neural network architecture, with a fully-connected structure and a Sequential model, which contains three layers: input layer, hidden layer and output layer.

This architecture was used for each different physical task. Thus, three neural networks were applied to estimate fatigue state from data of physical tasks (PA, SPI, MMH). All three networks have 40 nodes in the input layer, which represents the number of features extracted from the dataset, 16 nodes in the hidden layer and 4 nodes in the output layer (according to the 4 possible RPE values variations between each time window). In Fig. 2 the neural network architecture used for the three tasks is shown.

For all three neural networks, we use a ReLU (rectified linear) activation function in the input layer and in the hidden layer, on account of a better performance achieved with this function, and a Softmax activation function is applied to the output layer, which converts a vector of values to categorical probabilities. The category with more probability is chosen as the output in order to obtain a unique classification.

Regarding network's training parameters, we used the Adam gradient-based optimizer and the Categorical Crossentropy loss function, since the networks have to classify more than two classes. We first trained the neural networks with 500 epochs, and then with 1000 epochs, both processes with a batch size of 8. However, the three architectures employed differ in the learning rate used: for PA task, the learning rate is 0.01; for SPI task, 0.0001; and for MMH task, 0.005.

To adjust the final values of the hyperparameters (batch size and learning rate), 9 different tests were previously carried out for each task: three batch size and 3 learning rates were trained for each task. The values finally selected correspond to the option with the best result for each case. The tests carried out for this purpose are summarized in Table 1 for PA task, Table 2 for SPI task, and Table 3 for MMH task.

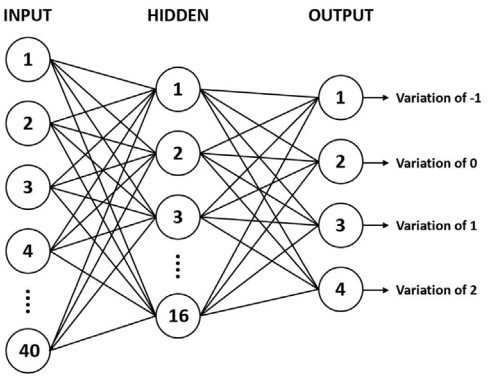

Fig. 2. Neural network architecture.

Table 1
Accuracy results after training data from PA task recordings with three different learning rates (LR) and three different batch sizes (BS).

PA task	LR 1e−4	LR 1e−3	LR 1e−2
BS 8	84.6%	94.2%	**98.1%**
BS 16	84.6%	90.4%	96.1%
BS 32	84.6%	88.5%	94.2%

Table 2
Accuracy results after training data from SPI task recordings with three different learning rates (LR) and three different batch sizes (BS).

SPI task	LR 1e−4	LR 1e−3	LR 1e−2
BS 8	**92.6%**	88.9%	87.0%
BS 16	90.7%	90.7%	87.0%
BS 32	88.9%	90.7%	77.8%

Table 3
Accuracy results after training data from MMH task recordings with three different learning rates (LR) and three different batch sizes (BS).

MMH task	LR .5e−4	LR .5e−3	LR .5e−2
BS 8	81.2%	87.5%	**93.7%**
BS 16	78.1%	84.3%	90.6%
BS 32	76.5%	85.9%	92.9%

2.3. Performance evaluation

At this point, the complete system used to classify the perceived exertion has been described. However, in order to evaluate the results obtained after the training process, we need to describe the metrics and mechanisms used for this task.

In total, three datasets have been created from the original data, in order to have one dataset for each physical task. Thus, the classification model was generated for each one of them, using 80% of each dataset for the neural network's training, and keeping 20% for testing purposes. Once the training is completed, the test phase of the model can be achieved by using the testing dataset with the final weights obtained from the training process. In order to make a complete performance evaluation, a set of metrics have been used for each class, which are detailed as follows:

$$accuracy = \sum_c \frac{TP_c + TN_c}{TP_c + TN_c + FP_c + FN_c}, c \in classes \tag{4}$$

$$sensitivity = \sum_c \frac{TP_c}{TP_c + FN_c}, c \in classes \tag{5}$$

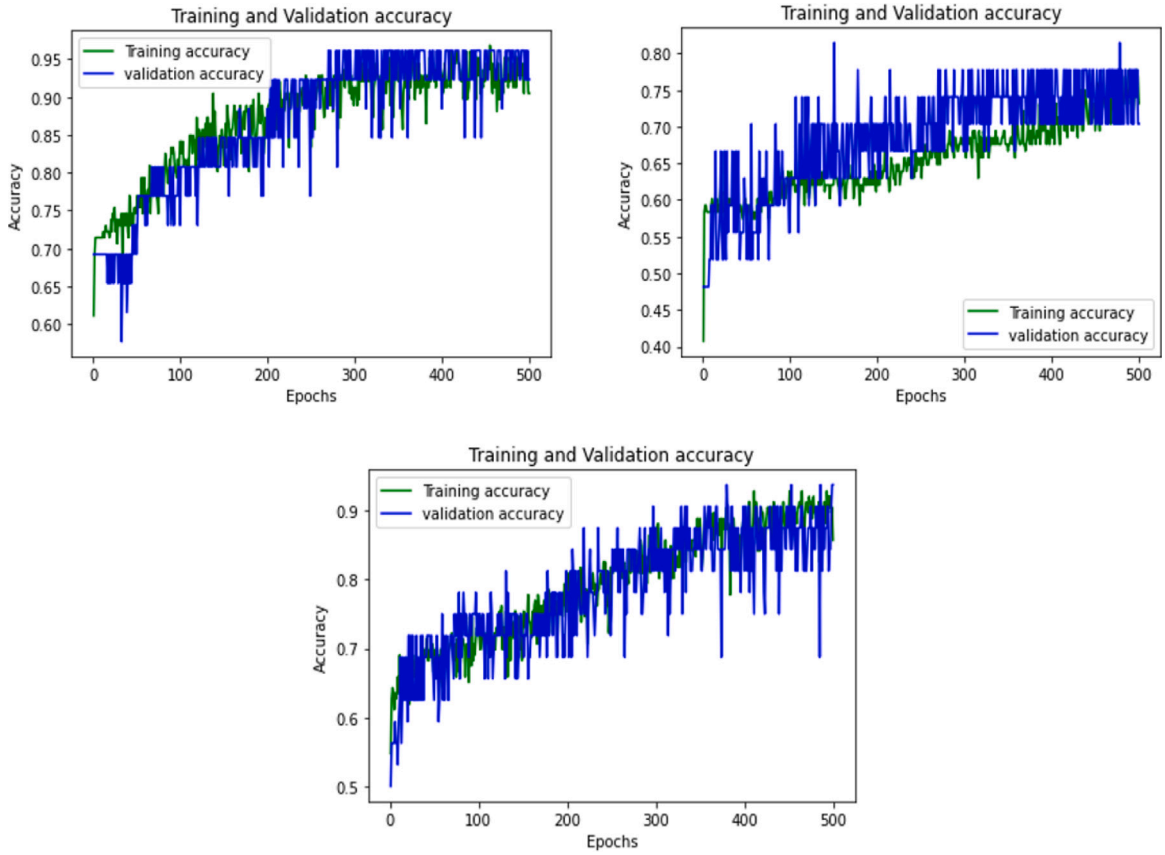

Fig. 3. Accuracy evolution during a 500-epoch training for PA task (top-left), SPI task (top-right) and MMH task (bottom).

Table 4
Total amount of samples obtained for each activity.

Activity	Train (80%)	Test (20%)	Total
PA	100 (10 for validation)	26	126
SPI	86 (9 for validation)	22	108
MMH	100 (10 for validation)	26	126

$$specificity = \sum_{c} \frac{TN_c}{TN_c + FP_c}, c \in classes \tag{6}$$

$$precision = \sum_{c} \frac{TP_c}{TP_c + FP_c}, c \in classes \tag{7}$$

$$F1_{score} = 2 \times \frac{precision \times sensitivity}{precision + sensitivity} \tag{8}$$

Where TP stands for True Positives, or the number of cases that belong to the class and are correctly classified; TN denotes True Negatives, or the number of cases that do not belong to the class and have not been classified as it; FP stands for False Positives, or the number of cases that do not belong to the class but are incorrectly classified as belonging to the class; and FN denotes False Negatives, which are the number of cases that belong to the class but are incorrectly classified as belonging to other class. In the following section, all of the performance results are indicated and compared.

2.3.1. Cumulative RPE evaluation

As previously indicated, the labelling and training of the system focuses on the variation of perceived exertion during each 10-minutes time window. With the evaluation described above, the performance of the system can be verified against these assumptions and the results will show how good the system is.

However, as an additional evaluation to this work, the cumulative perceived exertion for each user will be calculated independently based on the classification of our system in each time window and its accumulation until the completion of each 3-hour activity.

After this last study, we intend to evaluate the error that this type of cumulative systems based on time increments may have at the end of each complete activity. After these tests, the percentage of error compared with the final absolute labels of the dataset will be presented.

2.4. Related works

In order to verify the effectiveness of the proposed system, a comparison with similar works in recent years is ultimately carried out.

To this end, a search is carried out in the main search engines (Scopus, IEEExplorer and Google Scholar) with the following keywords: fatigue, machine learning, sensor (physiological or physical sensor). The results obtained are filtered by year, restricting the works to those published in the last 5 years (with the only restriction of the work published in 2013 by Zang et al. which is considered one of the first in the field and cited in most of the current works).

The results reflect a total of 8 papers after eliminating articles not focused on the design of a classifier and/or those with no citations. The selected works are presented below together with a brief summary of each one:

- (Zhang et al., 2013): binary muscle fatigue classifier using Support Vector Machines (SVM) with inertial sensors (IMU). Applied to walking sessions with 17 participants, looking the relationship between gait pattern and fatigue. 11 features selected, data normalization pre-processing and cross-validation technique applied in the testing phase. The classification is performed after the full trial.

Table 5
Metrics' results for PA task.

Epochs	*Class*	*TP*	*TN*	*FP*	*FN*	*Accuracy*	*Sensitivity*	*Specificity*	*Precision*	$F1_{score}$
500	−1	1	25	0	0	1	1	1	1	1
	0	18	6	2	0	0.923	1	0.75	0.9	0.947
	1	4	20	0	2	0.923	0.667	1	1	0.8
	2	1	25	0	0	1	1	1	1	1
	TOTAL	**24**	**76**	**2**	**2**	**0.961**	**0.923**	**0.974**	**0.923**	**0.923**
1000	−1	1	25	0	0	1	1	1	1	1
	0	18	8	0	0	1	1	1	1	1
	1	6	20	0	0	1	1	1	1	1
	2	1	25	0	0	1	1	1	1	1
	TOTAL	**26**	**78**	**0**	**0**	**1**	**1**	**1**	**1**	**1**

Table 6
Metrics' results for SPI task.

Epochs	*Class*	*TP*	*TN*	*FP*	*FN*	*Accuracy*	*Sensitivity*	*Specificity*	*Precision*	$F1_{score}$
500	−1	1	26	0	0	1	1	1	1	1
	0	13	10	4	0	0.852	1	0.714	0.765	0.867
	1	8	14	1	4	0.815	0.667	0.933	0.889	0.762
	2	1	26	0	0	1	1	1	1	1
	TOTAL	**23**	**76**	**5**	**4**	**0.917**	**0.852**	**0.938**	**0.821**	**0.836**
1000	−1	1	26	0	0	1	1	1	1	1
	0	13	11	3	0	0.889	1	0.785	0.812	0.896
	1	9	15	0	3	0.889	0.75	1	1	0.857
	2	1	22	0	0	1	1	1	1	1
	TOTAL	**24**	**74**	**3**	**3**	**0.942**	**0.889**	**0.961**	**0.889**	**0.889**

Table 7
Metrics' results for MMH task.

Epochs	*Class*	*TP*	*TN*	*FP*	*FN*	*Accuracy*	*Sensitivity*	*Specificity*	*Precision*	$F1_{score}$
500	−1	1	31	0	0	1	1	1	1	1
	0	16	12	2	2	0.875	0.889	0.857	0.889	0.889
	1	10	18	2	2	0.875	0.833	0.9	0.833	0.833
	2	1	31	0	0	1	1	1	1	1
	TOTAL	**28**	**92**	**4**	**4**	**0.937**	**0.875**	**0.958**	**0.875**	**0.875**
1000	−1	1	31	0	0	1	1	1	1	1
	0	17	14	0	1	0.968	0.944	1	1	0.971
	1	12	19	1	0	0.968	1	0.95	0.923	0.96
	2	1	31	0	0	1	**1**0.968	1	1	1
	TOTAL	**31**	**95**	**1**	**1**	**0.984**		**0.989**	**0.968**	**0.968**

- (Baghdadi et al., 2018): binary worker's fatigue classifier using SVM with an ankle-placed IMU. Applied to 3-hour manual material handling sessions with 14 participants. Kalman filtering pre-processing to extract motion features and detect kinematics. Classification based on the mean of distance-based scores (determined by the mean of the first 10 min fatigue level provided by participants, and the mean of the last 10 min fatigue level).
- (Karvekar et al., 2019): 2 and 4 class worker's fatigue classifier using SVM with ankle-placed smartphone. Experiment with 24 participants performing a fatiguing exercise (squatting), manually classifying the fatigue level after each 2-minutes exercise. Classification based on the gait pattern, using several features: mean, variation, acceleration, peaks, etc.
- (Nasirzadeh et al., 2020): several binary fatigue classifiers (k-nearest neighbours, Decision Tree, neural network, etc.) using heart rate signals, and manually classifying fatigue level after each 1-hour task. Experiments performed by 8 participants and consisted on three physical tasks (PA, SPI and MMH). Classification based on classic features like mean, variance and standard deviation, using cross-validation.
- (Sedighi Maman et al., 2017): fatigue classifier using several classifiers (random forest, SVM, Logistic regression, etc.) with four IMU sensors and one heart rate sensor. 24 participants performing MMH and SPI tasks. Classification based on several feature combinations (finally, less than 7 are used), using cross-validation.
- (Darbandy et al., 2020): physical fatigue classifier using KNN with heart rate signals. Experiments performed over 8 participants during 3-hour of MMH activities (divided in 1-hour periods). Classification based on nonlinear features (mean, standard deviation, minimum and maximum).
- (Lambay et al., 2021): binary fatigue classifier using recurrent neural networks (RNN) with IMU and heart rate signals. Using dataset with 18 participants performing manual tasks. Classification based on 23 features (not specified).
- (Kuschan & Krüger, 2021): 3 and 5 class worker's physical fatigue classifier using SVM with IMU sensors places in a exoskeleton. Classification based on 63 acceleration features.

As can be seen from the selected papers, there is no standard processing mechanism, as it depends on the sensors used and the type of classifier. The use of SVM and, curiously, the little use of neural networks stand out. Regarding the classified classes, many papers opt for defining a binary state (fatigue and non-fatigue) based on subdividing some scale of perceived exertion (such as RPE or CR10), while other papers extend the classes to define more ranges (as is the case of the last work). Among all the papers, only one of them classifies with the full range of values of the Borg scale (Lambay et al., 2021), but it will be seen in the results section that this has a negative impact on the classification.

With respect to the system proposed in this paper, which will be evaluated in the following section, it uses basic characteristics (mean, deviation, standard deviation, zero crossing, etc.) but extracted from

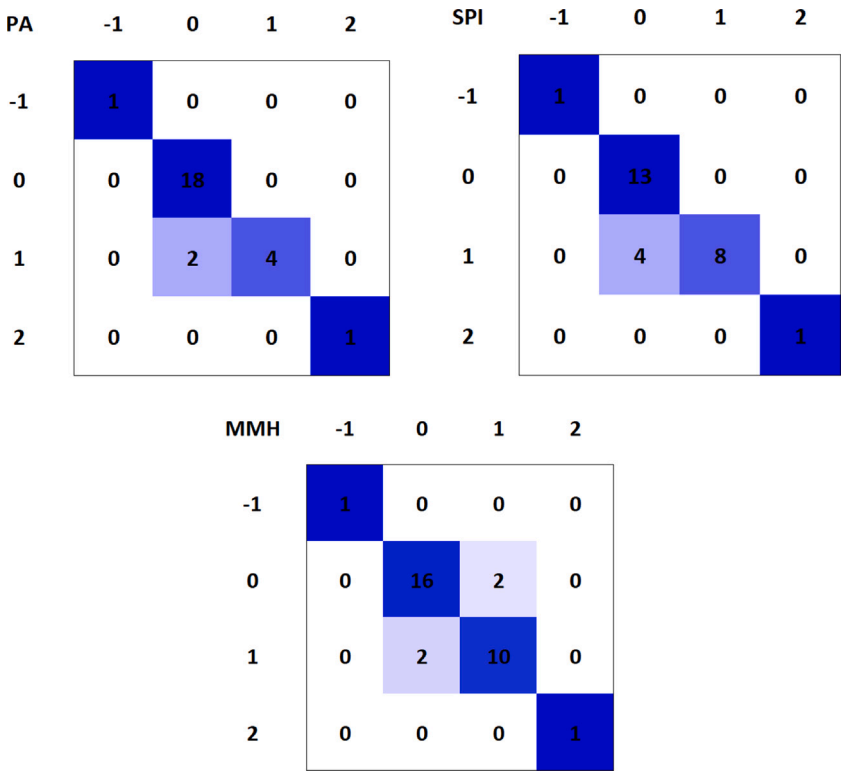

Fig. 4. Confusion matrix after 500 epochs for PA task (top-left), SPI task (top-right) and MMH task (bottom).

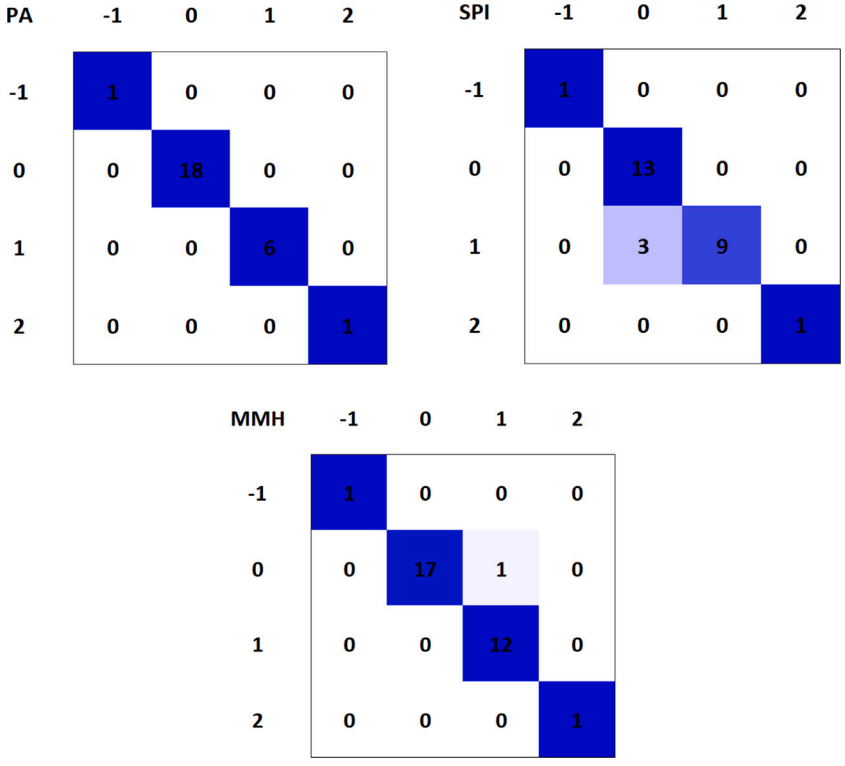

Fig. 5. Confusion matrix after 1000 epochs for PA task (top-left), SPI task (top-right) and MMH task (bottom).

the frequency components obtained from the application of the DWT. Among the previous works indicated, none can be found that uses this transform. Similarly, and as can be seen in this manuscript, this work classifies values belonging the full Borg scale (6 to 20), applying a novel mechanism for calculating absolute fatigue by means of the accumulation of fatigue variations in 10-minute intervals; this allows the number of classes to be classified to be reduced to 4, but maintaining the full range of the final scale (with which, this work presents a greater resolution in its detection).

Table 8
RPE for users 2–5 with 500-epoch network after each 10-min period (participant 5 did not perform SPI task). *Real* (*abs*): RPE from the questionnaires. *Real* (*var*): variation from previous checkpoint. *Classified*: network output (variation from previous checkpoint). *Acum*: accumulative estimation using the outputs of the network from the beginning to this point.

User	Task	Value	10'	20'	30'	40'	50'	1 h	1 h 10'	1 h 20'	1 h 30'	1 h 40'	1 h 50'	2 h	2 h 10'	2 h 20'	2 h 30'	2 h 40'	2 h 50'	3 h
2	PA	*Real (abs)*	7	8	9	9	9	9	9	10	10	10	10	10	11	11	11	11	11	12
		Real (var)	+1	+1	+1	+0	+0	+0	+0	+1	+0	+0	+0	+0	+1	+0	+0	+0	+0	+1
		Classified	+1	+1	+1	+0	+0	+0	+0	+1	+0	+0	+0	+0	+1	+0	+0	+0	+0	+1
		Acum	7	8	9	9	9	9	9	10	10	10	10	10	11	11	11	11	11	12
	SPI	*Real (abs)*	7	7	8	9	9	9	10	10	11	11	12	12	13	13	14	14	15	15
		Real (var)	+1	+0	+1	+1	+0	+0	+1	+0	+1	+0	+1	+0	+1	+0	+1	+0	+1	+0
		Classified	+1	+0	+1	+0	+0	+0	+1	+0	+1	+0	+1	+0	+1	+0	+1	+0	+0	+0
		Acum	7	7	8	8	8	8	9	9	10	10	11	11	12	12	13	13	13	13
	MMH	*Real (abs)*	7	8	8	9	9	9	10	10	11	11	11	12	12	12	13	14	15	15
		Real (var)	+1	+1	+0	+1	+0	+0	+1	+0	+1	+0	+0	+1	+0	+0	+1	+1	+1	+0
		Classified	+1	+1	+0	+1	+1	+0	+1	+0	+1	+0	+0	+1	+0	+0	+1	+1	+1	+0
		Acum	7	8	8	9	10	10	11	11	12	12	12	13	13	13	14	15	16	16
3	PA	*Real (abs)*	9	10	12	12	13	14	14	14	14	15	15	15	15	16	17	17	17	17
		Real (var)	+0	+1	+2	+0	+1	+1	+0	+0	+0	+1	+0	+0	+0	+1	+1	+0	+0	+0
		Classified	+0	+1	+2	+0	+1	+1	+0	+0	+0	+1	+0	+0	+0	+1	+1	+0	+0	+0
		Acum	9	10	12	12	13	14	14	14	14	15	15	15	15	16	17	17	17	17
	SPI	*Real (abs)*	7	9	11	12	12	13	13	13	14	15	16	16	17	17	17	18	18	19
		Real (var)	+0	+2	+2	+1	+0	+1	+0	+0	+1	+1	+1	+0	+1	+0	+0	+1	+0	+1
		Classified	+0	+2	+2	+1	+0	+1	+0	+0	+1	+0	+1	+0	+1	+0	+0	+1	+0	+1
		Acum	7	9	11	12	12	13	13	13	14	14	15	15	16	16	16	17	17	18
	MMH	*Real (abs)*	7	7	8	8	8	9	9	10	11	13	14	14	15	15	15	15	15	15
		Real (var)	+0	+0	+1	+0	+0	+1	+0	+1	+1	+2	+1	+0	+1	+0	+0	+0	+0	+0
		Classified	+0	+0	+1	+0	+0	+1	+0	+1	+1	+2	+1	+0	+1	+0	+0	+0	+0	+0
		Acum	7	7	8	8	8	9	9	10	11	13	14	14	15	15	15	15	15	15
4	PA	*Real (abs)*	7	7	8	10	10	10	10	10	10	12	12	12	12	12	13	13	13	13
		Real (var)	+0	+0	+1	+2	+0	+0	+0	+0	+0	+2	+0	+0	+0	+0	+1	+0	+0	+0
		Classified	+0	+0	+1	+2	+0	+0	+0	+0	+0	+2	+0	+0	+0	+0	+1	+0	+0	+0
		Acum	7	7	8	10	10	10	10	10	10	12	12	12	12	12	13	13	13	13
	SPI	*Real (abs)*	8	8	9	10	9	11	11	11	11	11	11	12	12	12	13	13	14	14
		Real (var)	+0	+0	+1	+1	-1	+2	+0	+0	+0	+0	+0	+1	+0	+0	+1	+0	+1	+0
		Classified	+0	+0	+1	+1	-1	+2	+0	+0	+0	+0	+0	+1	+0	+0	+1	+0	+1	+0
		Acum	8	8	9	10	9	11	11	11	11	11	11	12	12	12	13	13	14	14
	MMH	*Real (abs)*	9	10	11	11	11	11	11	11	12	12	13	14	14	14	14	14	14	14
		Real (var)	+0	+1	+1	+0	+0	+0	+0	+0	+1	+0	+1	+1	+0	+0	+0	+0	+0	+0
		Classified	+0	+1	+0	+0	+0	+0	+0	+0	+1	+0	+1	+1	+0	+0	+0	+0	+0	+0
		Acum	9	10	10	10	10	10	10	10	11	11	12	13	13	13	13	13	13	13
5	PA	*Real (abs)*	9	11	11	10	11	12	12	12	13	13	13	13	14	14	14	14	15	15
		Real (var)	+0	+2	+0	-1	+1	+1	+0	+0	+1	+0	+0	+0	+1	+0	+0	+0	+1	+0
		Classified	+0	+2	+0	-1	+1	+1	+0	+0	+1	+0	+0	+0	+1	+0	+0	+0	+1	+0
		Acum	9	11	11	10	11	12	12	12	13	13	13	13	14	14	14	14	15	15
	MMH	*Real (abs)*	11	12	12	12	12	12	13	13	13	13	13	14	14	14	15	14	15	15
		Real (var)	+0	+1	+0	+0	+0	+0	+1	+0	+0	+0	+0	+1	+0	+0	+1	-1	+1	+0
		Classified	+0	+1	+0	+0	+0	+0	+1	+0	+0	+0	+0	+1	+0	+0	+1	-1	+1	+0
		Acum	11	12	12	12	12	12	13	13	13	13	13	14	14	13	15	14	15	15

This comparison will be detailed deeply with quantitative results in the final part of the next section.

3. Results and discussion

In this section, the results obtained after training the neural network model with the three different physical tasks are shown, considering the metrics and the mechanism explained in the previous section, so that the goodness of the classification can be evaluated. After presenting the results, we examine different case studies for each task, to show that, after the prediction of the variation of the RPE values between 10-minutes time windows, the estimated RPE score between 6 and 20 can be calculated for the three-hours tasks as the accumulation of the variations. Comparing these case studies results with the actual values, we calculate the absolute error and show the goodness of the classification.

First, it is important to describe the size of each tasks' dataset used for the training and the testing phases. As explained in Section 2.1, the original dataset has been divided into 10-minutes time windows and the three different tasks have been grouped in three different datasets. Thus, from +800000 acceleration and jerk data samples, the final dataset was reduced to less than 400 samples with 40 features each. In Table 4 we show the division of the latter in three parts, mixed and randomized.

So, the content of the dataset is initially divided into a subset for training (80%) and a subset for testing (20%); but, in addition, to adjust the hyperparameters of the system (batch size and learning rate), the training set is also subdivided in two parts: one for training only and one for validation (10% of the initial training subset). The use of these three divisions is as follows:

- Training subset: used to find the correspondence between inputs and outputs of the network, adjusting the weights of the neurons' connections.
- Validation subset: used to validate the partial classification result of the network and serve as a goodness-of-fit aid for the real-time system in order to adjust hyperparameters.
- Testing subset: used to evaluate the final classification results. No neurons' connections weight adjustment can be performed at this point. The results obtained are used as the classifier final results.

However, as the division of these subsets is done randomly, slightly different results may be obtained in different training sessions. Therefore, cross-validation technique is applied to generate 10 random (and non-matching) combinations of training and test subsets. The results presented in the manuscript correspond to the arithmetic mean of the values obtained in these ten combinations.

Before checking the results it is important to mention that, after relabelling the samples in which we calculate the variation of the perceived exertion (instead of the absolute states), the four classes are not balanced. In fact, the classes with the most extreme values (−1 and 2) have few occurrences compared to the intermediate classes (0 and 1). This is due to the fact that, in 10-minute stretches for the tasks performed by the workers, the perceived exertions were not of very high or very low intensity, but of intermediate intensities. Therefore,

Table 9
RPE for users 6–8 with 500-epoch network after each 10-min period. *Real (abs)*: RPE from the questionnaires. *Real (var)*: variation from previous checkpoint. *Classified*: network output (variation from previous checkpoint). *Acum*: accumulative estimation using the outputs of the network from the beginning to this point.

User	Task	Value	10'	20'	30'	40'	50'	1 h	1 h 10'	1 h 20'	1 h 30'	1 h 40'	1 h 50'	2 h	2 h 10'	2 h 20'	2 h 30'	2 h 40'	2 h 50'	3 h
6	PA	*Real (abs)*	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7
		Real (var)	+0	+0	+0	+0	+0	+0	+0	+0	+1	+0	+0	+0	+0	+0	+0	+0	+0	+0
		Classified	+0	+0	+0	+0	+0	+0	+0	+0	+1	+0	+0	+0	+0	+0	+0	+0	+0	+0
		Acum	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7
	SPI	*Real (abs)*	9	11	13	14	14	14	14	15	15	15	15	15	15	16	16	17	17	18
		Real (var)	+0	+2	+2	+1	+0	+0	+0	+1	+0	+0	+0	+0	+0	+1	+0	+1	+0	+1
		Classified	+0	+2	+2	+1	+0	+0	+0	+1	+0	+0	+0	+0	+0	+0	+0	+0	+0	+0
		Acum	9	11	13	14	14	14	14	15	15	15	15	15	15	15	15	15	15	15
	MMH	*Real (abs)*	6	6	7	8	8	9	10	10	10	10	11	11	12	12	12	12	12	12
		Real (var)	+0	+0	+1	+1	+0	+1	+1	+0	+0	+0	+1	+0	+1	+0	+0	+0	+0	+0
		Classified	+0	+0	+7	+1	+1	+1	+1	+1	+0	+0	+1	+0	+1	+0	+0	+0	+1	+0
		Acum	6	6	7	8	9	10	11	12	12	12	13	13	14	14	14	14	15	15
7	PA	*Real (abs)*	6	6	7	7	8	8	9	9	9	10	10	10	11	12	12	12	13	14
		Real (var)	+0	+0	+1	+0	+1	+0	+1	+0	+0	+1	+0	+0	+1	+1	+0	+0	+1	+1
		Classified	+0	+0	+1	+0	+1	+0	+0	+0	+0	+1	+0	+0	+1	+1	+0	+0	+1	+1
		Acum	6	6	7	7	8	8	8	8	8	9	9	9	10	11	11	11	12	13
	SPI	*Real (abs)*	6	6	7	8	8	8	8	9	9	9	9	9	9	10	10	10	10	10
		Real (var)	+0	+0	+1	+1	+0	+0	+0	+1	+0	+0	+0	+0	+0	+1	+0	+0	+0	+0
		Classified	+0	+0	+1	+0	+0	+0	+0	+0	+0	+0	+0	+0	+0	+0	+0	+0	+0	+0
		Acum	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
	MMH	*Real (abs)*	6	7	7	7	7	8	8	9	10	10	10	11	11	11	12	12	12	12
		Real (var)	+0	+1	+0	+0	+0	+1	+0	+1	+1	+0	+0	+1	+0	+0	+1	+0	+0	+0
		Classified	+0	+1	+0	+0	+0	+0	+0	+1	+1	+0	+0	+1	+0	+0	+1	+0	+0	+0
		Acum	6	7	7	7	7	7	7	8	9	9	9	10	10	10	11	11	11	11
8	PA	*Real (abs)*	6	6	6	7	7	7	7	7	7	8	8	10	11	11	11	12	12	12
		Real (var)	+0	+0	+0	+1	+0	+0	+0	+0	+0	+1	+0	+2	+1	+0	+0	+1	+0	+0
		Classified	+0	+0	+0	+0	+0	+0	+0	+0	+0	+1	+0	+2	+1	+0	+0	+1	+0	+0
		Acum	6	6	6	6	6	6	6	6	6	7	7	9	10	10	10	11	11	11
	SPI	*Real (abs)*	8	8	8	9	11	12	12	13	13	13	14	14	14	14	14	15	14	14
		Real (var)	+0	+0	+0	+1	+2	+1	+1	+1	+0	+0	+1	+0	+0	+0	+0	+1	-1	+0
		Classified	+0	+0	+0	+1	+2	+1	+1	+1	+0	+0	+0	+0	+0	+0	+0	+1	-1	+0
		Acum	8	8	8	9	11	12	12	13	13	13	13	13	13	13	13	14	13	13
	MMH	*Real (abs)*	11	12	13	13	13	14	13	15	15	15	15	15	15	15	15	15	15	16
		Real (var)	+1	+1	+1	+0	+0	+1	-1	+2	+0	+0	+0	+0	+0	+0	+0	+0	+0	+1
		Classified	+1	+1	+1	+0	+0	+1	-1	+2	+0	+0	+0	+0	+0	+0	+0	+0	+0	+1
		Acum	11	12	13	13	13	14	13	15	15	15	15	15	15	15	15	15	15	16

in a 10-minute stretch for a physical activity of medium intensity, it is usual for the worker to experience qualitatively no increase in fatigue (class 0) or a slight increase (class 1).

As this is a problem caused by the dataset and the activities carried out, we have maintained the unbalancing of the classes, as it is adapted to a real case where the tasks involve a slight variation in fatigue in 10-minute stretches (if this were not the case, the worker would not be able to complete the working day satisfactorily).

After this clarification and using the parameters aforementioned for the neural network, we firstly apply a 500-epoch training and then use a random 20% of the dataset for testing and evaluate the validity of the model, with the metrics explained in Section 2.3. After this approach, we then applied the same process but using a 1000-epoch training for each of the three task dataset.

A quick summary about the accuracy tendency during the training process of each neural network (one for each activity) can be observed in Fig. 3.

The results obtained for PA task dataset with 500- and 1000-epoch training are shown in Table 5.

According to these metrics, for 500-epoch training, the overall error is very close to zero, as well as for each classification class. We can observe that classes −1 and 2 obtain a 100% accuracy because of the few occurrences that those classes have; but in the intermediate classes the accuracy results are very high too (more than 92%). However, for the class 1 the parameter *sensitivity* is low because of the relatively high amount of $false\ negatives$ with respect of the *true positives*; this metric does not affect the accuracy as the amount of *truenegatives* is very high (and that situation mitigate the problems caused by the $false\ negatives$). However, the final metrics obtained in this case show acceptable results: 96.1% *accuracy*, 92.3% *sensitivity*, 97.4% $specificity$, 92.3% *precision* and 92.3% $F1_{score}$.

To solve the problems detected before for the class 1, the neural network is trained with 1000 epochs. For this case, results improve significantly, obtaining a perfect classification without any sample misclassified. So all the evaluation metrics obtain a value of 100%.

The same processing is performed for the network related to SPI task dataset. The results obtained are presented in Table 6.

For the 500-epoch training, results are similar to the ones obtained for PA task. We can observe that class 1 obtains a *sensitivity* results very low, because of the amount of $false\ negatives$ classified. For this task, as there are more misclassifications than the previous one, more errors are observed for the intermediate classes, obtaining acceptable but lower results for the metrics used to evaluate the system: 91.7% *accuracy*, 85.2% *sensitivity*, 93.8% $specificity$, 82.1% *precision* and 83.6% $F1_{score}$.

However, as observed before, the 1000-epoch training obtains a more accurate neural network. For this task, there is not a perfect classification (as observed for PA task), but the results improve significantly. The results obtained increase around 5% in all metrics with this second network.

Finally, following the same process, evaluation metrics for MMH task dataset are exposed in Table 7.

In this case, there is a difference: as can be observed, MMH task requires more time to be performed (or has be performed during more time) because this dataset has more samples than the others. However, the behaviour around the classes distribution is the same: the most populated classes are 0 and 1, but there are more samples on these classes. Working with a larger dataset allow us to better generalize the system behaviour. In fact, as there are more samples, the $false\ negative$ cases are reduced in a significant way; so the *sensitivity* parameter has a higher value for this task. However, as detailed in the previous cases, performing a 1000-epoch training improves the results obtained as can be seen in Table 7.

For all three task datasets, it can be observed that evaluation metrics are worse for classes of 0 and 1, while classes of −1 and 2 obtain an error of zero for all metrics. That case was commented at the beginning, but it can be summarized that it is due to the fact that the number of

Table 10
RPE for users 2–5 with 1000-epoch network after each 10-min period (participant 5 did not perform SPI task). *Real (abs)*: RPE from the questionnaires. *Real (var)*: variation from previous checkpoint. *Classified*: network output (variation from previous checkpoint). *Acum*: accumulative estimation using the outputs of the network from the beginning to this point.

User	Task	Value	10'	20'	30'	40'	50'	1 h	1 h 10'	1 h 20'	1 h 30'	1 h 40'	1 h 50'	2 h	2 h 10'	2 h 20'	2 h 30'	2 h 40'	2 h 50'	3 h
2	PA	*Real (abs)*	7	8	9	9	9	9	9	10	10	10	10	10	11	11	11	11	11	12
		Real (var)	+1	+1	+1	+0	+0	+0	+0	+1	+0	+0	+0	+0	+1	+0	+0	+0	+0	+1
		Classified	+1	+1	+1	+0	+0	+0	+0	+1	+0	+0	+0	+0	+1	+0	+0	+0	+0	+1
		Acum	7	8	9	9	9	9	9	10	10	10	10	10	11	11	11	11	11	12
	SPI	*Real (abs)*	7	7	8	9	9	9	10	10	11	11	12	12	13	13	14	14	15	15
		Real (var)	+1	+0	+1	+1	+0	+0	+1	+0	+1	+0	+1	+0	+1	+0	+1	+0	+1	+0
		Classified	+1	+0	+1	+0	+0	+0	+1	+0	+1	+0	+1	+0	+1	+0	+1	+0	+1	+0
		Acum	7	7	8	8	8	8	9	9	10	10	11	11	12	12	13	13	14	14
	MMH	*Real (abs)*	7	8	8	9	9	9	10	10	11	11	11	12	12	12	13	14	15	15
		Real (var)	+1	+1	+0	+1	+0	+0	+1	+0	+1	+0	+0	+1	+0	+0	+1	+1	+1	+0
		Classified	+1	+1	+0	+1	+0	+0	+1	+0	+1	+0	+0	+1	+0	+0	+1	+1	+1	+0
		Acum	7	8	8	9	9	9	10	10	11	11	11	12	12	12	13	15	15	15
3	PA	*Real (abs)*	9	10	12	12	13	14	14	14	14	15	15	15	15	16	17	17	17	17
		Real (var)	+0	+1	+2	+0	+1	+1	+0	+0	+0	+1	+0	+0	+0	+1	+1	+0	+0	+0
		Classified	+0	+1	+2	+0	+1	+1	+0	+0	+0	+1	+0	+0	+0	+1	+1	+0	+0	+0
		Acum	9	10	12	12	13	14	14	14	14	15	15	15	15	16	17	17	17	17
	SPI	*Real (abs)*	7	9	11	12	12	13	13	13	14	15	16	16	17	17	17	18	18	19
		Real (var)	+0	+2	+2	+1	+0	+1	+0	+0	+1	+1	+1	+0	+1	+0	+0	+1	+0	+1
		Classified	+0	+2	+2	+1	+0	+1	+0	+0	+1	+0	+0	+0	+0	+0	+1	+1	+1	+1
		Acum	7	9	11	12	12	13	13	13	14	14	14	14	14	14	15	16	17	18
	MMH	*Real (abs)*	7	7	8	8	8	9	9	10	11	13	14	14	15	15	15	15	15	15
		Real (var)	+0	+0	+1	+0	+0	+1	+0	+1	+1	+2	+1	+0	+1	+0	+0	+0	+0	+0
		Classified	+0	+0	+1	+0	+0	+1	+0	+1	+1	+2	+1	+0	+1	+0	+0	+0	+0	+0
		Acum	7	7	8	8	8	9	9	10	11	13	14	14	15	15	15	15	15	15
4	PA	*Real (abs)*	7	7	8	10	10	10	10	10	10	12	12	12	12	12	13	13	13	13
		Real (var)	+0	+0	+1	+2	+0	+0	+0	+0	+0	+2	+0	+0	+0	+0	+1	+0	+0	+0
		Classified	+0	+0	+1	+2	+0	+0	+0	+0	+0	+2	+0	+0	+0	+0	+1	+0	+0	+0
		Acum	7	7	8	10	10	10	10	10	10	12	12	12	12	12	13	13	13	13
	SPI	*Real (abs)*	8	8	9	10	9	11	11	11	11	11	11	12	12	12	13	13	14	14
		Real (var)	+0	+0	+1	+1	-1	+2	+0	+0	+0	+0	+0	+1	+0	+0	+1	+0	+1	+0
		Classified	+0	+0	+1	+1	-1	+2	+0	+0	+0	+0	+0	+1	+0	+0	+1	+0	+1	+0
		Acum	8	8	9	10	9	11	11	11	11	11	11	12	12	12	13	13	14	14
	MMH	*Real (abs)*	9	10	11	11	11	11	11	11	12	12	13	14	14	14	14	14	14	14
		Real (var)	+0	+1	+1	+0	+0	+0	+0	+0	+1	+0	+1	+1	+0	+0	+0	+0	+0	+0
		Classified	+0	+1	+1	+0	+0	+0	+0	+0	+1	+0	+1	+1	+0	+0	+0	+0	+0	+0
		Acum	9	10	11	11	11	11	11	11	12	12	13	14	14	14	14	14	14	14
5	PA	*Real (abs)*	9	11	11	10	11	12	12	12	13	13	13	13	14	14	14	14	15	15
		Real (var)	+0	+2	+0	-1	+1	+1	+0	+0	+1	+0	+0	+0	+1	+0	+0	+0	+1	+0
		Classified	+0	+2	+0	-1	+1	+1	+0	+0	+1	+0	+0	+0	+1	+0	+0	+0	+1	+0
		Acum	9	11	11	10	11	12	12	12	13	13	13	13	14	14	14	14	15	15
	MMH	*Real (abs)*	11	12	12	12	12	12	13	13	13	13	13	14	14	14	15	14	15	15
		Real (var)	+0	+1	+0	+0	+0	+0	+1	+0	+0	+0	+0	+1	+0	+0	+1	-1	+1	+0
		Classified	+0	+1	+0	+0	+0	+0	+1	+0	+0	+0	+0	+1	+0	+0	+1	-1	+1	+0
		Acum	11	12	12	12	12	12	13	13	13	13	13	14	14	13	15	14	15	15

samples for the first and last classes are much less than for the classes of 0 and 1, since it is less likely that the physical fatigue decreases in one during effort, or that it increases in two in such a short time as 10 min.

In order to evaluate the performance of the perceived exertion model, confusion matrices for all three tasks for the case of 500-epoch training are shown in Fig. 4 for the three tasks.

With the confusion matrices, we can appreciate that the few errors occur between the medium values of 0 and 1, which indicates that the error is acceptable, since the model does not classify a 0 value as 2 or vice versa. Following the same steps, the confusion matrices are shown for all three test datasets for a 1000-epoch training in Fig. 5 for the three tasks.

As expected, results are better with a 1000-epoch training than the ones obtained with a 500-epoch training, even for PA dataset, where there are no errors; although, for SPI dataset three mistaken classification occur from estimating a variation label of 1 as 0, and one classification error takes place for MMH dataset predicting a variation of 0 as 1.

As a final aspect to analyse the system proposed in this work, and relating the results to the study of perceived exertion during the full working day, an exhaustive study has been carried out comparing the accumulation of the partial classifications of effort for each worker independently. In this way, at the end of the working day, we are able to compare the variation of absolute perceived exertion (by means of partial accumulations of ratings every 10 min) with respect to the absolute values labelled in the original dataset.

This study, although it does not provide an additional classification or additional information on the system, helps us to observe qualitatively the variation of the cumulative model with respect to the labelled model. We recall at this point that, in previous works, this dataset has been used to classify on the basis of absolute values but, according to the results obtained, there is no direct relationship; this is why the cumulative model, after the results shown above, shows that it is the ideal one for this type of system. The study was performed to all the workers, but due to errors in the recording process, participant 1 is not included. Results are shown in Table 8 (users 2–5) and Table 9 (users 6–8) for the 500-epoch training; and in Table 10 (users 2–5) and Table 11 (users 6–8) for the 1000-epoch training. After this evaluation, the summary of the results obtained for all the workers is presented.

As can be observed in Tall those tables, the first two rows of each activity indicate the real RPE value obtained from the questionnaires after each 10-minute evaluation, and the RPE variation between one time period to another. The next rows are the results related to the classification system: the third one is the output of the classifier (RPE variation) and the fourth is the accumulate RPE from the beginning (that cam be compared with the real RPE obtained from the questionnaires). Values coloured in green are those whose values (RPE variation and RPE accumulated) are exactly the same as the real ones; values in red are those related with misclassifications of the neural network (RPE variation classified does not correspond to real RPE variation); and values in yellow indicate those classifications where the classifier hit but, because of a previous miss the RPE accumulated is not the same as the real one. It is important to mention that, for the neural network, only those values coloured in red are harm the accuracy results.

Table 11
RPE for users 6–8 with 1000-epoch network after each 10-min period. *Real (abs)*: RPE from the questionnaires. *Real (var)*: variation from previous checkpoint. *Classified*: network output (variation from previous checkpoint). *Acum*: accumulative estimation using the outputs of the network from the beginning to this point.

User	Task	Value	10'	20'	30'	40'	50'	1 h	1 h 10'	1 h 20'	1 h 30'	1 h 40'	1 h 50'	2 h	2 h 10'	2 h 20'	2 h 30'	2 h 40'	2 h 50'	3 h
6	PA	*Real (abs)*	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7
		Real (var)	+0	+0	+0	+0	+0	+0	+0	+0	+1	+0	+0	+0	+0	+0	+0	+0	+0	+0
		Classified	+0	+0	+0	+0	+0	+0	+0	+0	+1	+0	+0	+0	+0	+0	+0	+0	+0	+0
		Acum	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7
	SPI	*Real (abs)*	9	11	13	14	14	14	14	15	15	15	15	15	15	16	16	17	17	18
		Real (var)	+0	+2	+2	+1	+0	+0	+0	+1	+0	+0	+0	+0	+0	+1	+0	+1	+0	+1
		Classified	+0	+2	+2	+1	+0	+0	+0	+1	+0	+0	+0	+0	+0	+1	+0	+1	+0	+1
		Acum	9	11	13	14	14	14	14	15	15	15	15	15	15	16	16	17	17	18
	MMH	*Real (abs)*	6	6	7	8	8	9	10	10	10	10	11	11	12	12	12	12	12	12
		Real (var)	+0	+0	+1	+1	+0	+1	+1	+0	+0	+0	+1	+0	+1	+0	+0	+0	+0	+0
		Classified	+0	+0	+7	+1	+0	+1	+1	+0	+0	+0	+1	+0	+1	+0	+0	+0	+0	+0
		Acum	6	6	7	8	8	9	10	10	10	10	11	11	12	12	12	12	12	12
7	PA	*Real (abs)*	6	6	7	7	8	8	9	9	9	10	10	10	11	12	12	12	13	14
		Real (var)	+0	+0	+1	+0	+1	+0	+1	+0	+0	+1	+0	+0	+1	+1	+0	+0	+1	+1
		Classified	+0	+0	+1	+0	+1	+0	+1	+0	+0	+1	+0	+0	+1	+1	+0	+0	+1	+1
		Acum	6	6	7	7	8	8	9	9	9	10	10	10	11	12	12	12	13	14
	SPI	*Real (abs)*	6	6	7	8	8	8	8	9	9	9	9	9	9	10	10	10	10	10
		Real (var)	+0	+0	+1	+1	+0	+0	+0	+1	+0	+0	+0	+0	+0	+1	+0	+0	+0	+0
		Classified	+0	+0	+1	+1	+0	+0	+0	+1	+0	+0	+0	+0	+0	+1	+0	+0	+0	+0
		Acum	6	6	7	8	8	8	8	9	9	9	9	9	9	9	10	10	10	10
	MMH	*Real (abs)*	6	7	7	7	7	8	8	9	10	10	10	11	11	11	12	12	12	12
		Real (var)	+0	+1	+0	+0	+0	+1	+0	+1	+1	+0	+0	+1	+0	+0	+1	+0	+0	+0
		Classified	+0	+1	+0	+0	+0	+1	+0	+1	+1	+0	+0	+1	+0	+0	+1	+0	+0	+0
		Acum	6	7	7	7	7	8	8	9	10	10	10	11	11	11	12	12	12	12
8	PA	*Real (abs)*	6	6	6	7	7	7	7	7	7	8	8	10	11	11	11	12	12	12
		Real (var)	+0	+0	+0	+1	+0	+0	+0	+0	+0	+1	+0	+2	+1	+0	+0	+1	+0	+0
		Classified	+0	+0	+0	+1	+0	+0	+0	+0	+0	+1	+0	+2	+1	+0	+0	+1	+0	+0
		Acum	6	6	6	7	7	7	7	7	7	8	8	10	11	11	11	12	12	12
	SPI	*Real (abs)*	8	8	8	9	11	12	12	13	13	13	14	14	14	14	14	15	14	14
		Real (var)	+0	+0	+0	+1	+2	+1	+1	+1	+0	+0	+1	+0	+0	+0	+0	+1	-1	+0
		Classified	+0	+0	+0	+1	+2	+1	+1	+0	+0	+0	+1	+0	+0	+0	+0	+1	-1	+0
		Acum	8	8	8	9	11	12	12	12	12	12	13	13	13	13	13	14	13	13
	MMH	*Real (abs)*	11	12	13	13	13	14	13	15	15	15	15	15	15	15	15	15	15	16
		Real (var)	+1	+1	+1	+0	+0	+1	-1	+2	+0	+0	+0	+0	+0	+0	+0	+0	+0	+1
		Classified	+1	+1	+1	+0	+0	+1	-1	+2	+0	+0	+0	+0	+0	+0	+0	+0	+0	+1
		Acum	11	12	13	13	13	14	13	15	15	15	15	15	15	15	15	15	15	16

Table 12
Accuracy comparison with other classifiers using the same preprocessing chain and the same extracted features from the dataset.

Accuracy	SVM	RNN	NN (500-epoch)	NN (1K-epoch)
PA task	81%	92.3%	96.1%	100%
SPI task	77%	74.1%	91.7%	94.2%
MMH task	66%	82.8%	93.7%	98.4%

If we take into account only the absolute RPE obtained after the complete work activity, nine of the twenty cases presented in Tables 8 and 9 obtain a 0% error, seven obtain a 5% error (as the Borg scale maximum value is 20), one obtains a 10% error and three obtain a 15% error. If we check the results for all the participants of the dataset, the mean error obtained in the absolute RPE value at the end of every work task is 4.25%.

The same study was performed with the full dataset using the neural network trained after 1000 epochs. As could be observed in the results tables and the confusion matrices, the results improve. However, the same is study for users 2 to 8 is presented in Tables 10 and 11. As can be observed in those tables, misclassifications are reduced in all the users (with the exception of user 3). For the 1000-epoch trained network, seventeen of the twenty cases presented in Tables 10 and 11 obtain a 0% error and the other three obtain a 5% error (only fail in one classification). If we check the final results for all the participants, the mean error obtained in the absolute RPE value at the end of every work task is less than 1% (0.75%).

With these two accumulative studies we can affirm that the classification system based on RPE variations every 10-minute window is suitable for estimating absolute RPE with an error less than 1% for the best system and 4.25% for the worst system.

However, classical neural networks are not the only technology that allows the design of such classifiers. As we will see in the next comparison with previous works, there are multiple statistical and/or artificial intelligence mechanisms that can help in this task. To corroborate that the mechanism used in this work is indeed the most adequate, the same preprocessing and feature extraction has been applied to two other classifiers: a simpler classifier (SVM) and a more complex one (RNN). The summary of the results can be seen in Table 12.

Finally, after presenting and analysing the results of the RPE estimation system to detect physical fatigue, we show a comparison in Table 13 with similar studies related to the prediction of physical effort in work environments. The works presented have some differences between them: sensors used, tasks performed and collected in the dataset used, technology used for the classification system and metric classified. However, there are not very much works similar to this in the area, and that is the reason some of them have important differences.

As can be observed in Table 13, the first improvement of this work compared with last years' studies is the accuracy obtained with the neural network model. Although the main difference between our system and the others is the partial RPE classification as the variation between one period to the next one, we have experimentally verified that this classification can be used to estimate the absolute RPE value (calculated as the accumulate value of the predicted variation between each time window). So, our system obtains the best accuracy results from all of the works included in the comparison. Moreover, the results presented in Table 13 are the ones regarding the 500-epoch training neural network; so, using the second implementation presented in this work, the results would be much better.

On the other hand, in other works with a high accuracy (Baghdadi et al., 2018; Karvekar et al., 2019; Zhang et al., 2013), the classification system is trained to detect only binary fatigue (no fatigue vs. fatigue), which does not give such a precise fatigue result. Other investigations have predicted Borg's RPE values from 6 to 20; however, as the number of classes to classify increases, the accuracy gets worse. That are the

Table 13
Comparison with recent works.

Study	Year	Sensors	Tasks	Technology	Detection	Accuracy
Zhang et al.	2013	IMU	Squatting Walking	SVM	Binary fatigue	96%
Baghdadi et al.	2018	IMU	MMH	SVM	Binary fatigue	90%
Karvekar et al.	2019	IMU	Walking	SVM	Binary fatigue 4 classes	91% 65%
Nasirzadeh et al.	2020	HR	PA, SPI, MMH	NN KNN NB DT RF RI LiR LoR LDA	RPE	90.4%, 69.4%, 77.8% 86.6%, 60.5%, 73.7% 79.9%, 51.1%, 64.7% 85.3%, 58.4%, 72.9% 86.6%, 60.4%, 75.9% 87.8%, 71.9%, 79.7% 86.2%, 64.7%, 77.3% 82.8%, 59.2%, 77.7% 87.6%, 66.0%, 77.3%
Sedighi Maman et al.	2020	IMU HR	SPI, MMH	RF BAG BO SVM LoR PLoR	RPE	89.7%, 87.9% 88.6%, 87.0% 88.0%, 87.0% 78.7%, 82.0% 62.4%, 77.8% 80.0%, 85.9%
Darbandy et al.	2020	HR	MMH	KNN	RPE	78.2%
Lambay et al.	2021	IMU HR	SPI, MMH	RNN	Binary fatigue	65%
Kuschan and Krüger	2021	IMU	PA	SVM	3-class CR10 5-class CR10	93.3% 83.8%
This work (500-epoch)	2021	IMU	PA, SPI, MMH	NN	RPE	96.1%, 91.7%, 93.7%
This work (1000-epoch)	2021	IMU	PA, SPI, MMH	NN	RPE	100%, 94.2%, 98.4%

Legend:
IMU: Inertial Movement Unit
PA: Part Assembly
SVM: Support Vector Machine
NB: Naive Bayes
RI: Rule Induction
LDA: Linear Discriminant Analysis
PLoR: Penalized Logistic Regression
RPE: Rating of Perceived Exertion
HR: Heart Rate
SPI: Supply Pickup and Insertion
NN: Neural Network
DT: Decision Tree
LiR: Linear Regression
BAG: Bagging
RNN: Recurrent Neural Network
CR10: Normalized 0–10 Borg Scale
MMH: Manual Material Handling
KNN: k-Nearest Neighbours
RF: Random Forest
LoR: Logistic Regression
BO: Boosting

cases of works (Darbandy et al., 2020; Nasirzadeh et al., 2020; Sedighi Maman et al., 2020).

And, finally, there are other works that try to classify the RPE absolute value (Lambay et al., 2021). However, the experience obtained after this work shows that the calculation of the absolute RPE value through partial accumulations is more effective and better results are obtained.

If we analyse the similarities of our system with the others, we can find that the most common activities proposed to the participants are PA, SPI and MMH; but not all the papers use all the activities: the most used is MMH, and few are the papers that include the PA activity. Furthermore, with regard to the subjective scale of perceived fatigue, the Borg scale (in its classic version from 6 to 20, or in its standardized version CR10 from 0 to 10) is the most widely used. However, because the scale encompasses multiple values, it is common for studies to reduce the number of classes (grouping values close together) in order to obtain better results. Finally, regarding the sensors used to measure activity, our work uses the most common sensor (IMU), as it is the sensor whose response is most closely linked to physical activity. It is true that heart rate also varies, but it is a subject-dependent variable (a fitter worker would have less variation in heart rate than a less fit worker, even if both were doing the same activity).

Finally, if we focus on the differences found between this work and previous works, it is worth highlighting the use of classical neural networks (which are only used in one previous work) and the inclusion of features extracted from frequency components (only one previous work uses frequency components, but it uses the Fourier transform, while this work applies the DWT). Last but not least, the calculation of the final fatigue as the accumulation of fatigue variations over the day is only used in this work. Other works make intermediate measurements, but provide absolute (not relative) values.

With the approach proposed in this study about predicting the RPE variation values between 10-minute time windows, we obtain less classes, which gives a higher accuracy than studies that predicted absolute RPE values. Furthermore, as we calculate RPE values from the predicted variations, we can obtain a better precision than those works that only detected between fatigue and non-fatigue state.

4. Conclusions

In this work, the importance of monitoring the fatigue state of a user is presented and analysed. This task is of particular interest to workers who perform physical tasks that cause heavy physical wear and tear.

Using a public dataset containing information from several workers performing three different tasks on several occasions during temporary periods of 3 h duration, a Machine Learning system has been designed, implemented and tested for the automatic classification of the variations of the perceived exertion during the time slots labelled in the dataset (of 10 min duration).

For this purpose, a frequency study of the variation of the different sensors within the time slots has been carried out, extracting the most relevant features that are used as inputs of the neural network.

The classification results for 500 and 1000 epochs training show positive results in the accuracy of the system (over 91% in all cases), which demonstrates that this system is viable for use as a detector of the worker's state of fatigue.

In order to obtain the absolute fatigue results, the system has been evaluated by accumulating the partial classification results of each time slot, in order to obtain the final absolute fatigue value for each user and compare it with the labelled values of the dataset. The results show an error of less than 3% in all cases, supporting the feasibility of the system.

If we compare this work with other similar works in the field of worker fatigue classification, the results show an improvement in all cases. Moreover, in other works, fatigue levels are reduced to improve the results or classification mechanisms of the absolute value of fatigue (not relative) are used; but, even so, the results of this work are better compared to them.

Finally, it can be concluded that a relationship between physical activity and fatigue using the information provided by inertial sensors has been established, demonstrating the usefulness of this system. For future work, this system will be integrated in an embedded system to provide real-time recommendations; and, on the other hand, it will be adapted to study the fatigue in pathologies that cause a sudden increase in physical fatigue (such as chronic fatigue syndrome).

CRediT authorship contribution statement

Elena Escobar-Linero: Software, Tests, Writing. **Manuel Domínguez-Morales:** Conceptualization, Methodology, Writing. **José Luis Sevillano:** Conceptualization, Writing, Funding.

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.

Acknowledgements

This work has been supported by "Fondo Europeo de Desarrollo Regional" (FEDER) and "Consejería de Economía, Conocimiento, Empresas y Universidad" of the Junta de Andalucía, under Programa Operativo FEDER 2014–2020 (Project US-1263715).

References

Aghdam, S., et al. (2019). Fatigue assessment scales: A comprehensive literature review. *Archives of Hygiene Sciences*, *8*, 145–153.

Ahamed, N. U., et al. (2021). Using machine learning and wearable inertial sensor data for the classification of fractal gait patterns in women and men during load carriage. *Procedia Computer Science*, *185*, 282–291.

Al-Saegh, A., et al. (2021). Deep learning for motor imagery EEG-based classification: A review. *Biomedical Signal Processing and Control*, *63*, Article 102172.

Al-shair, K., et al. (2016). The effect of fatigue and fatigue intensity on exercise tolerance in moderate COPD. *Lung*, *194*(6), 889–895.

Atiya, S. O., et al. (2021). Accelerometer-based physical fatigue assessment in 400 meter running event. *AIP Conference Proceedings*, *2339*(1), Article 020188.

Ayrulu-Erdem, B., & Barshan, B. (2011). Leg motion classification with artificial neural networks using wavelet-based features of gyroscope signals. *Sensors*, *11*, 1721–1743.

Baghdadi, A., et al. (2018). A machine learning approach to detect changes in gait parameters following a fatiguing occupational task. *Ergonomics*, *61*(8), 1116–1129.

Baghdadi, A., et al. (2019). Monitoring worker fatigue using wearable devices: A case study to detect changes in gait parameters. *Journal of Quality Technology*, *53*(1), 47–71.

Balkin, T. J., et al. (2011). The challenges and opportunities of technological approaches to fatigue management. *Accident Analysis and Prevention*, *43*(2), 565–572, Advancing Fatigue and Safety Research.

Barim, M. S., et al. (2019). Accuracy of an algorithm using motion data of five wearable IMU sensors for estimating lifting duration and lifting risk factors. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 1105–1111.

Battini, D., et al. (2014). Innovative real-time system to integrate ergonomic evaluations into warehouse design and management. *Computers & Industrial Engineering*, *77*, 1–10.

Beniczky, S., et al. (2021). Machine learning and wearable devices of the future. *Epilepsia*, *62*(S2), S116–S124.

Borg, G. (1982). Psychophysical bases of perceived exertion. *Medicine and Science in Sports and Exercise*, *14*(5), 377—381.

Borg, G. (1990). Psychophysical scaling with applications in physical work and the perception of exertion. *Scandinavian Journal of Work, Environment & Health*, *16*, 55–58.

Borg, E. (2007). *On perceived exertion and its measurement* (Ph.D. thesis), Stockholm University.

Braccesi, C., et al. (2015). Random multiaxial fatigue: A comparative analysis among selected frequency and time domain fatigue evaluation methods. *International Journal of Fatigue*, *74*, 107–118.

Buckley, C., et al. (2017). Binary classification of running fatigue using a single inertial measurement unit. In *2017 IEEE 14th international conference on wearable and implantable body sensor networks* (pp. 197–201).

Bunevicius, A., et al. (2011). Relationship of fatigue and exercise capacity with emotional and physical state in patients with coronary artery disease admitted for rehabilitation program. *American Heart Journal*, *162*(2), 310–316.

Busch, A. J., et al. (2011). Exercise therapy for fibromyalgia. *Current Pain and Headache Reports*, *15*(5), 358–367.

Chowdhury, S. K., et al. (2013). Discrete wavelet transform analysis of surface electromyography for the fatigue assessment of neck and shoulder muscles. *Journal of Electromyography and Kinesiology*, *23*(5), 995–1003.

Darbandy, M., et al. (2020). A new approach to detect the physical fatigue utilizing heart rate signals. *Research in Cardiovascular Medicine*, *9*, 23.

Dominguez-Morales, M., et al. (2019). Smart footwear insole for recognition of foot pronation and supination using neural networks. *Applied Sciences*, *9*, 3970.

Dong, H., et al. (2014). Development of a fatigue-tracking system for monitoring human body movement. In *2014 IEEE international instrumentation and measurement technology conference (I2MTC) proceedings* (pp. 786–791).

Dzeng, R.-J., et al. (2014). A feasibility study of using smartphone built-in accelerometers to detect fall portents. *Automation in Construction*, *38*, 74–86.

European Commission (2010). *Health and safety at work in Europe*.

Fu, J., et al. (2019). Continuous measurement of muscle fatigue using wearable sensors during light manual operations. In *Digital human modeling and applications in health, safety, ergonomics and risk management. Human body and motion* (pp. 266–277). Springer International Publishing.

Gamberale, F. (1985). The perception of exertion. *Ergonomics*, *28*(1), 299–308.

Gawron, V., French, J., & Funke, D. (2001). An overview of fatigue..

Goldenberg, M. M. (2012). Multiple sclerosis review. *Pharmacy and Therapeutics*, *37*(3), 175–184.

Greenberg, S., & Frid, M. (2006). Chronic fatigue syndrome–exercise and physical activity. *Harefuah*, *145*(4), 276–80, 318.

Health and Safety Executive (2006). Managing shift work: Health and safety guidance.

Jiang, Y., et al. (2021). A data-driven approach to predict fatigue in exercise based on motion data from wearable sensors or force plate. *Sensors*, *21*(4).

Karvekar, S., et al. (2019). A data-driven model to identify fatigue level based on the motion data from a smartphone. *BioRxiv*, http://dx.doi.org/10.1101/796854.

Karvekar, S., et al. (2021). Smartphone-based human fatigue level detection using machine learning approaches. *Ergonomics*, *64*(5), 600–612.

Krupp, L. B., et al. (1989). The fatigue severity scale: Application to patients with multiple sclerosis and systemic lupus erythematosus. *Archives of Neurology*, *46*(10), 1121–1123.

Kuschan, J., & Krüger, J. (2021). Fatigue recognition in overhead assembly based on a soft robotic exosuit for worker assistance. *CIRP Annals*, *70*(1), 9–12.

Lambay, A., et al. (2021). A data-driven fatigue prediction using recurrent neural networks. In *2021 3rd International congress on human-computer interaction, optimization and robotic applications* (pp. 1–6). IEEE.

Lamooki, S. R., et al. (2020). Challenges and opportunities for statistical monitoring of gait cycle acceleration observed from IMU data for fatigue detection. In *2020 8th IEEE RAS/EMBS international conference for biomedical robotics and biomechatronics* (pp. 593–598). IEEE.

Lang, B., et al. (1981). Autoimmune aetiology for myasthenic (eaton-lambert) syndrome. *Lancet (London, England)*, *2*(8240), 224–226.

Lee, S., & Lee, D. (2020). Healthcare wearable devices: An analysis of key factors for continuous use intention. *Service Business*, *14*, 1–29.

Lih, O. S., et al. (2020). Comprehensive electrocardiographic diagnosis based on deep learning. *Artificial Intelligence in Medicine*, *103*, Article 101789.

Luna-Perejón, F., et al. (2019). Wearable fall detector using recurrent neural networks. *Sensors*, *19*(22), 4885.
Luna-Perejón, F., et al. (2021). Ankfall—Falls, falling risks and daily-life activities dataset with an ankle-placed accelerometer and training using recurrent neural networks. *Sensors*, *21*(5), 1889.
Masala, D., et al. (2017). Physical activity and its importance in the workplace. *Igiene E Sanita Pubblica*, *73*(2), 159–169.
Mitchell, E., et al. (2013). Classification of sporting activities using smartphone accelerometers. *Sensors*, *13*, 5317–5337.
Mostafa, H., et al. (2017). Wearable devices in medical internet of things: Scientific research and commercially available devices. *Hir*, *23*(1), 4–15.
Muñoz-Saavedra, L., et al. (2020). Affective state assistant for helping users with cognition disabilities using neural networks. *Electronics*, *9*(11).
Nasirzadeh, F., et al. (2020). Physical fatigue detection using entropy analysis of heart rate signals. *Sustainability*, *12*(7).
National Safety Council (2020). *Fatigue: Developing an effective fatigue risk management system*.
O'Connor, L., et al. (2020). Myasthenia gravis and physical exercise: A novel paradigm. *Frontiers in Neurology*, *11*.
Ricci, J., et al. (2007). Fatigue in the U.S. workforce: Prevalence and implications for lost productive work time. *Journal of Occupational and Environmental Medicine / American College of Occupational and Environmental Medicine*, *49*, 1–10.
Ritchie, C. (2012). Rating of perceived exertion (RPE). *Journal of Physiotherapy*, *58*(1), 62.
Sadeghniiat, K., & Yazdi, Z. (2015). Fatigue management in the workplace. *Industrial Psychiatry Journal*, *24*, 12–17.
Schmidt, M., et al. (2016). IMU-based determination of fatigue during long sprint. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct* (pp. 899–903).
Sedighi Maman, Z., et al. (2017). A data-driven approach to modeling physical fatigue in the workplace using wearable sensors. *Applied Ergonomics*, *65*.
Sedighi Maman, Z., et al. (2020). A data analytic framework for physical fatigue management using wearable sensors. *Expert Systems with Applications*, *155*, Article 113405.
Sekine, M., et al. (2000). Classification of waist-acceleration signals in a continuous walking record. *Medical Engineering & Physics*, *22*(4), 285–291.
Seneviratne, S., Hu, Y., Nguyen, T., Lan, G., Khalifa, S., Thilakarathna, K., et al. (2017). A survey of wearable devices and challenges. *IEEE Communications Surveys & Tutorials*, *19*(4), 2573–2620.
Strohrmann, C., et al. (2012). Monitoring kinematic changes with fatigue in running using body-worn sensors. *IEEE Transactions on Information Technology in Biomedicine : A Publication of the IEEE Engineering in Medicine and Biology Society*, *16*, 983–990.
Tee, K. S., et al. (2017). A study on the ergonomic assessment in the workplace. *AIP Conference Proceedings*, *1883*(1), Article 020034.
Vignais, N., et al. (2013). Innovative system for real-time ergonomic feedback in industrial manufacturing. *Applied Ergonomics*, *44*(4), 566–574.
Vøllestad, N. K. (1997). Measurement of human muscle fatigue. *Journal of Neuroscience Methods*, *74*(2), 219–227.
Watt, B., & Grove, R. (1993). Perceived exertion: Antecedents and applications. *Sports Medicine*, *15*, 225–241.
Yung, M., et al. (2014). Detecting within- and between-day manifestations of neuromuscular fatigue at work: An exploratory study. *Ergonomics*, *57*(10), 1562–1573.
Zhang, J., et al. (2013). Classifying lower extremity muscle fatigue during walking using machine learning and inertial sensors. *Annals of Biomedical Engineering*, *42*.
Zhang, L., et al. (2019). Automated monitoring of physical fatigue using jerk.
Zhang, J., et al. (2020). Automatic detection of dynamic and static activities of the older adults using a wearable sensor and support vector machines. *Sci*, *2*(3).