Computers & Industrial Engineering 188 (2024) 109859

Contents lists available at ScienceDirect

Computers & Industrial Engineering

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

Physiological workload assessment for highly flexible fine-motory assembly tasks using machine learning

Markus Brillinger [a,c,*], Samuel Manfredi [a], Dominik Leder [a], Martin Bloder [c], Markus Jäger [b], Konrad Diwold [a], Amer Kajmakovic [a], Michael Haslgrübler [b], Rudolf Pichler [c], Martin Brunner [d], Stefan Mehr [e], Viktorijo Malisa [f]

[a] Pro2Future GmbH, Inffeldgasse 25F, 8010 Graz, Austria
[b] Pro2Future GmbH, Altenberger Strasse 69, 4040 Linz, Austria
[c] Institute of Production Engineering, Graz University of Technology, Inffeldgasse 25F, 8010 Graz, Austria
[d] Antemo GmbH, Gewerbepark 6, 8755 St. Peter ob Judenburg, Austria
[e] sanSirro GmbH, Stangersdorf-Gewerbegebiet 110, 8403 Lebring, Austria
[f] AUVA, Wienerbergstraße 11, 1100 Wien, Austria

ARTICLE INFO

Keywords:
Assembly of small-volume products
Commercially available wearable low-cost sensor
Random forest and K-Nearest-Neighbours
Workload assessment

ABSTRACT

In assembly of small-volume products, tasks are still frequently executed manually. However, the lead times foreseen for these tasks, often do not take into account the actual capabilities of the employees, which in turn leads to increased workload and the associated stress among the employees. This paper investigates how a commercially available wearable low-cost sensor and two machine learning algorithms can be applied to measure and evaluate heart rate, heart rate variability and respiration rate to establish a relationship with workload. The investigated algorithms, namely Random Forest and K-Nearest-Neighbours are able to distinguish between tasks phases and rest phases as well as between easy and difficult tasks executed by the employee, which is the main novelty of this paper.

1. Introduction

The automation of assembly tasks for high-volume products has been part of the industrial standard for a long time and is also becoming increasingly important for small-volume products (Calawa & Smith, 2017; Johansen, Rao, & Ashourpour, 2021). However, the assembly of small-volume products is often not profitable, which is why many activities are still executed manually (Kalscheuer, Eschen, & Schüppstuhl, 2021). The lead time for these manual assembly tasks is often based on the industrial used method of time measurement (MTM) (Breznik, Buchmeister, & Vujica Herzog, 2023; Laring, Forsman, Kadefors, & Örtengren, 2002). However, if the MTM lead times exceed the individual capabilities of the employees, the increased workload leads to stress, which manifests itself in increased sick leave, increased accidents at work and higher error rates in assembly (Báez, Rodriguez, Limon, & Tlapa, 2014; Kern & Refflinghaus, 2015; Saptari, Leau, & Mohamad, 2015; Thorvald, Lindblom, & Andreasson, 2019). However, this counteracts securing the long-term utilisability of the employee in assembly. This paper contributes to mitigating this gap between MTM lead time and individual human capabilities by developing a minimally invasive solution that can be applied during assembly tasks to avoid stress to the employees.

2. State of the art

2.1. Assembly line optimiziation

The human factor plays a crucial role in assembly line planning and optimisation, but there is still a gap between the state of research and the industrial praxis (Boysen, Fliedner, & Scholl, 2008). The biggest challenge is the variance in the individual capabilities of assembly employees since their capabilities depend on a variety of factors, such as motivation, the working environment or (mental and physical) stress (Tempelmeier, 2003). The latter factor in particular is examined in the context of assembly processes to create more precise assembly line models for optimisation: Studies in electronic industries indicate that self-reported stress assessment in parallel with heart rate variability (HRV) measurement is a physiological marker of stress (Mahmad Khairai, Abdul Wahab, & Sutarto, 2022). In the

* Corresponding author.
E-mail address: markus.brillinger@pro2future.at (M. Brillinger).
URL: https://www.pro2future.at (M. Brillinger).

https://doi.org/10.1016/j.cie.2023.109859
Received 3 May 2023; Received in revised form 20 November 2023; Accepted 23 December 2023
Available online 3 January 2024

automotive industry, heart rate, oxygen consumption and subjective evaluations during screw-driving operations confirm that heavy loads, side bending and twisting postures are very harmful to employees from both a physiological and biomechanical perspective (Chung, Lee, & Yeo, 2001). More complex studies for assembly tasks using biomarkers point out that self-reports and physiological vital parameters like catecholamine and cortisol responses are associated selectively with different psychological conditions: Catecholamine values are associated with feelings of time pressure and pressure by demands, cortisol values with irritation, tension and tiredness (Lundberg, Granqvist, Hansson, Magnusson, & Wallin, 1989). Considering mental fatigue in assembly line optimisation, experiments prove that the multi-objective optimisation method is particularly accurate in the area of static or relatively slow manual handling operations in assembly (Ma, Zhang, Chablat, Bennis, & Guillaume, 2009). Regarding employees' rest allowance and smoothing of the workload, Finco et al. propose a method, that allows optimisation of the performance of the assembly process considering not only productivity aspects but also the employee's well-being in assembly processes (Finco, Battini, Delorme, Persona, & Sgarbossa, 2020). Melin et al. examine how the organisation of assembly work affects stress based on systolic blood pressure, heart rate and adrenaline: Two different ways of organising assembly work are compared, a more traditional assembly line with fixed workstations as a chain and with short, repetitive work cycles and a new and more flexible work organisation with small autonomous groups. The latter have greater opportunities to influence the pace and content of their work and have thereby been able to significantly reduce stress (Melin, Lundberg, Söderlund, & Granqvist, 1999). From this one can conclude, that the type of assembly task has a significant influence on the employees' perceived psychological workload and physiological stress response.

2.2. Workload assessment

"Workload" is a hypothetical construct developed within the domain of human factors (HF) psychology. In it, various workload measurement techniques are used to evaluate equipment or workplaces in terms of the workload experienced by people using them. This workload construct emerged from extensive, task-specific research on the capacities and limitations of the human information processing system (MacDonald, 2003). The workload is defined as an objective measure of the demand of the work. This includes the complexity and the amount of tasks to execute (Hagmüller, Rank, & Kubin, 2006). To classify the workload, two categories of methods are the standard:

The subjective assessment methods assume that a human can assess and evaluate one's workload (Sweller, Van Merrienboer, & Paas, 1998). The main part of these methods is the use of a questionnaire in which the human rate the level of workload. Commonly used assessment techniques are the NASA-Task Load Index (NASA-TLX), Subjective Workload Dominance Technique (SWORD) (Stanton, Salmon, Walker, Baber, & Jenkins, 2005) and the Subjective Workload Assessment Technique (SWAT) (Rubio, Díaz, Martín, & Puente, 2004; Thorvald et al., 2019). The NASA-TLX score is interpreted with a given workload scale: low (0–9), medium (10–29), somewhat high (30–49), high (50–79), and very high (80–100).

The task/performance-based assessment methods of workload are intended to provide an objective measurement independent of individual factors. These focus on the actual tasks and the time spent executing these tasks. One example of this approach is the Method Time Measurement (MTM).

2.3. Stress assessment

A general consensus on the definition of stress has not been formulated since the nature of stress and its perception and therefore the invoked human coping responses are very dependent on the individual (Alsuraykh, Wilson, Tennent, & Sharples, 2019). One definition of stress on which this paper is based declares stress as a non-equivalent measure between the workload imposed on a person and the ability to cope with that workload (Alsuraykh et al., 2019). Stress in assembly tasks comes from many different forms of workload, for example, working while standing, working with heavy loads, working in a kneeling position or overhead. Stress can also have different effects on employees. Persons felt stress physiologically (e.g., acceleration of heartbeat and breathing), emotionally (e.g., frustration, anxiety, feelings of fatigue), and behaviourally (e.g., concentration problems, increase in errors) (Baua, 2020). Various techniques exist to assess workload-induced stress and its physiological responses, divided into two major domains: subjective workload assessment and physiological stress response measurement. For the latter one, several physiological stress response indicators represent workload-induced stress (Sweller et al., 1998). These indicators include heart activity, specifically the heart rate and the heart rate variability, brain activity, eye activity, skin conductance/response, bio-markers, such as cortisol, and SpO2 saturation:

2.3.1. Heart activity

Detecting the physiological stress response of the cardiovascular system via heart rate (HR), respiration rate (RR) and heart rate variability (HRV) measured with an electrocardiogram (ECG) empower to observe the response of the cardiovascular system to external stimuli like workload (Solange et al., 1981). Putting employees in a situation with a high workload, the relative changes of HR, RR and HRV indicate the cardiovascular stress responses of humans (Goldberger, Goldberger, & Shvilkin, 2018). The continuous development and improvement of mobile sensors for HRV monitoring has reduced the cost of equipment and the required application effort for these systems.

2.3.2. Brain activity

Electroencephalography (EEG) is used to measure brain activity in, for example, human–computer interaction to measure the physiological stress response to complete tasks (Kumar & Kumar, 2016; Kumar & kumar, 2016). EEG findings are obtained by recording electrical voltage fluctuations across the skin provoked by neuronal activity within the cortex (He, Mahfouf, & Torres-Salomao, 2018). To measure these voltage fluctuations, neuro-headsets with multiple channels are used to record comprehensive data that needs extensive analysis by experts to draw conclusions (Kumar & Kumar, 2016).

2.3.3. Eye activity

To assess physiological stress response without contact and without excessive restriction of movement, Fridman et al. use video-based data acquisition. Cameras measure and analyse the movements of the pupil (eye movement) and eyelid (eye blinking). The results of experimental measurements performed on air traffic controllers. Ahlstrom and Friedman-Berg (2006) indicate, that the time spent blinking the eyes is shorter when a high workload is applied to the subject. This paper highlighted that the mean pupil diameter is significantly larger for subjects under high workload (Engström, Markkula, Victor, & Merat, 2017; Kun et al., 2011).

2.3.4. Skin conductivity

Galvanic Skin Response (GSR), also known as Electrodermal Activity (EDA), is an easily acquired, non-intrusive and physiological signal used to measure physiological stress response. With this inexpensive and robust measurement method, electrical conductance is measured by sensors on the skin, at the foot or hand (Mehler, Reimer, Coughlin, & Dusek, 2009). The sweating of the human body changes the conductance of the skin due to the altered moisture on the skin. These body responses can be attributed to the nervous system. However, these physiological stress responses are highly individual and cannot be used as the only physiological stress response indicator without extensive individual data analysis and subjective assessment methods (Nourbakhsh,

Wang, Chen, & Calvo, 2012). Further research indicates, that it is possible to determine a stressful situation by only observing GSR data but it requires a lot of individual data analysis (Bakker, Pechenizkiy, & Sidorova, 2011).

2.3.5. Biomarkers

Biomarkers are molecules found in the body fluids of humans. These molecules quantify physiological stress response and provide information about the condition and health of the body (Carrasco & Van de Kar, 2003). Cortisol is one of these biomarkers and is part of saliva, blood, cerebrospinal fluid, urine and sweat. For measuring the physiological stress response of humans, fluids such as saliva or sweat are mainly studied in more detail because they are easy to sample and reliable (Samson & Koh, 2020). To acquire the data, electrochemical measurement methods are often used which convert the biochemical signals into electrical signals by using electrodes (Cho, Kim, & Park, 2020).

2.3.6. SpO2 saturation

The oxygen saturation (SpO2) in the human bloodstream represents the ratio of oxygen-saturated haemoglobin compared to unsaturated haemoglobin. In a healthy adult, this ratio varies between 97% and 100% (Bachner, 2003). The SpO2 as a physiological stress response is often used in combination with other indicators such as cardiovascular markers or body temperature (Akmandor & Jha, 2017). On the other hand tissue oxygen saturation (StO2) proves to be a reliable non-invasive stress response indicator (Chen, Yuen, Richardson, Liu, & She, 2014). However, the measurement of the StO2 parameters occurs via a visual system where the test subject is centered in the focus of the camera hence reducing free mobility during manual task execution.

3. Research gap

In the presented state of the art, a wide variety of experiments are conducted to detect the physiological stress responses. However, these experiments are executed exclusively under laboratory conditions with highly accurate and expensive measuring equipment accompanied by specialised personnel, which limits the use of these sensors in the industry. As a consequence, the authors of this paper derive the following research questions: Is it possible to recognise and distinguish physiological stress responses caused by different workloads in assembly tasks using (i) a commercially available wearable low-cost sensor that acquire heart rate, heart rate variability and respiration rate data, and utilise (ii) standard machine learning algorithms such as Random Forrest (RF) and K-Nearest-Neighbours (KNN) to differentiate between the stress responses induced by different workloads?

4. Approach

To answer the research question, the authors of this paper investigate a commercially available wearable low-cost sensor for data acquisition of physiological stress response (HR, HRV, RR) and utilise 2 machine learning algorithms (KNN, RF) for data processing and compare the results with the subjective assessment method (NASA-TLX).

For this purpose, experiments are conducted for assembly tasks. The physiological stress response is measured by the relative changes in HRV and HR. In addition, a subjective assessment of workload is conducted using the NASA-TLX method. The results of both methods are compared afterwards. In the experiments, no special considerations are given to baseline measurements since the expected results should be independent of individual baselines. To investigate the robustness of the two machine learning algorithms, cross-validation for a possible generalised stress differentiation model is performed.

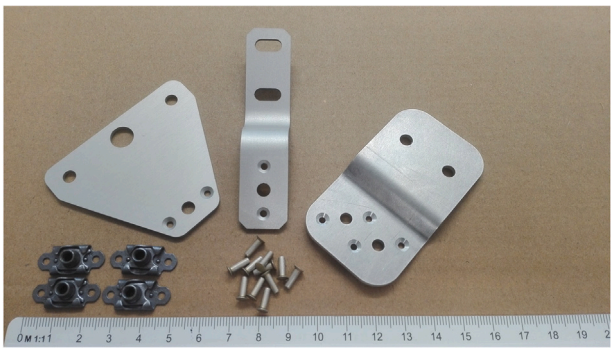
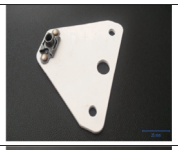
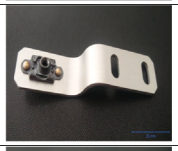


Fig. 1. The parts for assembling the module: base plates (silver colour), nut plates (dark colour) and countersunk head rivets (brassy colour).

Table 1
Modules, parts, and MTM-based lead times.

Module	Figure	Parts	MTM-Time
A		1 base plate, 1 nut plate, 2 rivets	9s
B		1 base plate, 1 nut plate, 2 rivets	9s
C		1 base plate, 2 nut plates, 4 rivets	15s

4.1. Assembly tasks

Three different modules of the aerospace industry are defined as the basis of the experiments. These modules, that have to be assembled, differ in shape and number of individual parts, but hardly in size and weight (all $< 18g$), depicted in Fig. 1.

All three modules consist of a base plate, nut plates and countersunk head rivets. First, one or two nut plates are inserted into an insertion plate, then the base plate is positioned on top and finally, up to 4 rivets (two for each nut plate) are inserted into the countersunk holes.

Due to the good fit as well as the small size and weight of the parts, physical stress, as caused by lifting bigger weights or the application of large forces, is largely eliminated in this setup. This reduces the assembly task to a challenge that mostly requires fine-motor skills and therefore a high degree of concentration by the subject. To induce workload-related physiological stress to the subject, the assembly tasks must be performed by each subject faster than calculated by the Method Time Measurement (MTM) (see Table 1).

The assembly tasks are of two kinds: easy and difficult. For an easy task, the subjects are asked to perform only the last steps of the assembly process for the module (e.g. inserting the rivets) whereas for a difficult task, the subjects have to assemble the whole module within the same amount of time as for the easy task.

4.2. Workplace

The workplace is equipment with a monitor, a mouse and a keyboard, as well as the insertion plate for the parts and all the materials

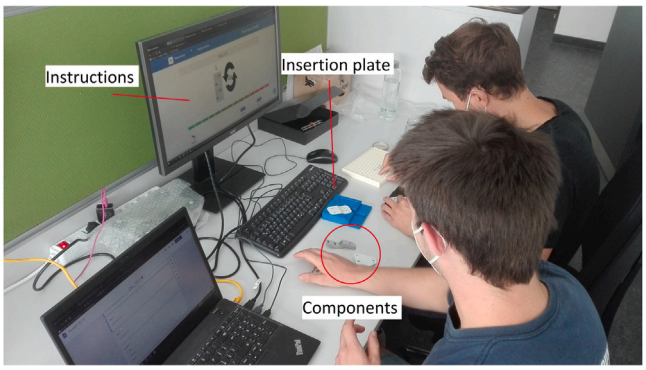


Fig. 2. Structure of the workplace for the tests.

Fig. 3. Sensor shirt and multifunctional sensor module used in this paper.

needed to perform the assembly task. The monitor displays the work instructions, as well as a quantitative time display in the form of a coloured segmented bar. The assembly task instructions consist of an image with associated textual instructions, as represented in Fig. 2. Throughout the experiments, the subjects sit on a chair to avoid any excessive physical workload.

4.3. Sensors and signals

To measure the physiological stress response of a subject, a commercially available wearable low-cost sensor is used which detects HR, HRV and RR. The acquisition of GSR and EEG data, as suggested in the state of the art, is not possible. The sensor consists of a sensor shirt and a multifunctional sensor module, depicted in Fig. 3. The cost of this sensor is currently 398,- EUR.

Using this sensor has advantages for possible deployment in an industrial setting compared to specialised equipment used for GSR or biomarkers. Furthermore, the sensor does not require specialised personnel to use and the costs for this sensor are low compared to medical ECG or eye-tracking devices. Last, the subjects are not restricted in their freedom of movement as it would be by having single electrodes applied with the medical ECG or EEG.

The measuring principle of the sensor is based on the dry electrode measuring method with electrolytic half-cell potentials (Chi, Jung, & Cauwenberghs, 2010; Ramasamy & Balan, 2018). The half-cell potentials are connected together with series resistances between the skin and a dry electrode. The dry electrode is made of conductive silicone. With this, the voltage signals of the heartbeat are detected and transmitted to the multifunctional sensor module via electrical conductors, which are incorporated into the sensor shirt. In this way, the data is acquired with sampling intervals of 100 ms (HRV, RR) and 500 ms (HR).

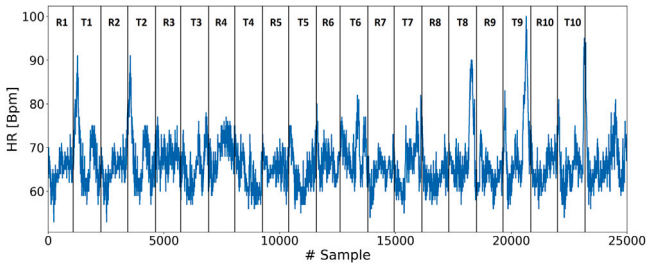


Fig. 4. Raw HR data of subject 0 at day 0, distinguished in assembly task (indicated as T) and rest phase (indicated as R).

4.4. Subjects

A total of six exclusively male healthy subjects in an age group between 25 and 38 years are selected to eliminate gender-, age- and education-related variations in the results. Written consent is obtained from each subject to process their physiological stress response data, which constitutes highly personal data. However, due to the small number of subjects and the associated low statistical power of the experiments, it is not possible to draw conclusions about heterogeneous groups of subjects. However, this is not the aim of this paper.

4.5. Test execution

The experiments are conducted exclusively in the morning, on two different days. The subjects have never tried or seen the tasks before. Before the experiments start, the subjects are equipped with the sensor to acquire the data for physiological stress response (HR, HRV and RR).

On the first day, the experiments are divided into the rest phase and the easy task phase. The easy task phase can be interpreted as the phase where workload is induced. Each phase has a 2-minute duration. Both phases alternate ten times for a total experiment duration of 40 min. The sequence of the phases is $R_1 - T_1 - R_2 - T_2 - \cdots - T_{10}$, where R_i declares a rest phase and T_i declares a easy task phase. During the rest phases, the subjects are asked to relax on the chair.

On the second day, the subjects must execute the difficult task. These circumstances should put the subjects in a situation where the subject's impression is that their capabilities are not adequate to complete the given task within the lead time and therefore induce an increased workload. To increase the workload during the task phase, the work instructions are accompanied by a down-counting timer for the completion of the assembly tasks. After each completed assembly task, the subject must click on the mouse to record the actual assembly time required.

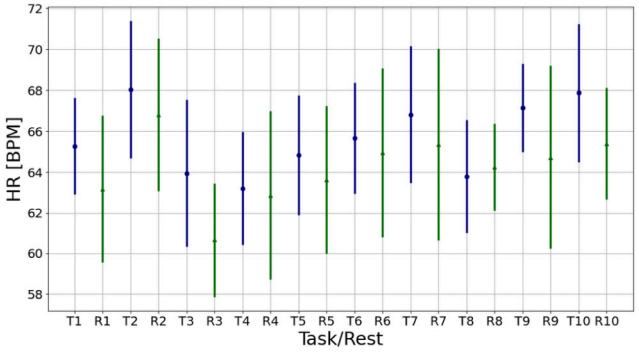
In both experiments, the subjects are asked to complete a subjective workload assessment based on NASA-TLX halfway through the experiment. The timing of the workload assessment is based on the fact that after the experiment a learning curve is to be expected and therefore the workload might seem reduced during the experiment.

5. Data processing

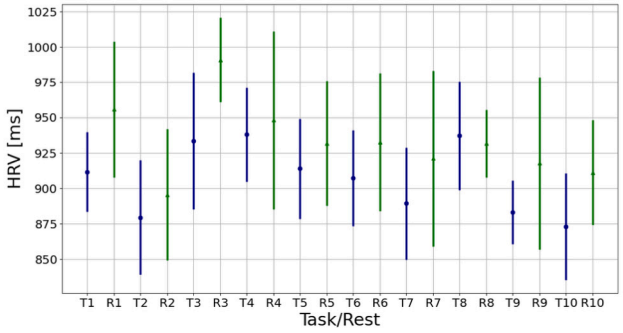
5.1. Raw data and feature selection

Fig. 4 illustrates the raw HR data acquired during the experiments. All data acquisition devices were synchronised on Unix time before the experiments started. No additional data pre-processing was necessary in order to proceed with the data analysis.

For data processing and evaluating, the mean value of the HR and HRV are the major features as suggested by the related research. Furthermore, the RR is used but exclusively for the training of the RF and the KNN.



(a) HR mean value and standard deviation over total experiment



(b) HRV mean value and standard deviation over total experiment

Fig. 5. Differentiation between rest phases and assembly tasks during the experiments. Blue lines and dots represent the assembly task phase mean value and standard deviation. Green lines and dots represent the rest phase mean value and standard deviation.

5.2. Data analysis

The first question that arises is whether HR and HRV of different subjects have the same underlying distribution which is the null hypothesis. The Wilcoxon Signed-Rank Test (Rey & Neuhäuser, 2011) as well as the Mann–Whitney U-Test (Ramachandran & Tsokos, 2015) indicate a rejection of this null hypothesis ($\alpha = 0.05$, $p = 0$). Therefore one can conclude that the physiological stress response data have significant differences in the underlying distributions.

To examine the relative changes of the mean value and standard deviation of the HR and HRV in each rest and task phase, Fig. 5 indicates, that the mean value of the HRV decreases while the HR increases during a task phase when compared to the subsequent rest phase. This observation is consistent with the findings of other researchers (Caroline Chanel, Wilson, & Scannella, 2019; Fahr & Hofer, 2013; Fridman, Reimer, Mehler, & Freeman, 2018). Hence, one can conclude that the sensor is capable of recognising changes in the HR and HRV due to workload-induced physiological stress responses. However, only by visually analysing features extracted from the two experiments, there is no significant visual difference between the data from the easy and the difficult task phase. Therefore, machine learning methods can provide deeper insights and test whether a mathematical model can distinguish between a difficult and an easy task phase.

Fig. 6 presents the variance of HR and HRV mean values during task phases. For instance, the mean value of the HR during rest phase 0 is approx. 80 BPM whereas the mean value of the HR during rest phase 8 is approx. 75 BPM. The same holds for the HRV mean value at rest phase 0 with 750 ms and the HRV mean value at rest phase 8 with 790 ms.

A further comparison is done in terms of the mean value of HR and HRV in rest phases as illustrated in Fig. 7: The mean value of HR during the rest phases is 80 BPM on the first day and 69 BPM on the second

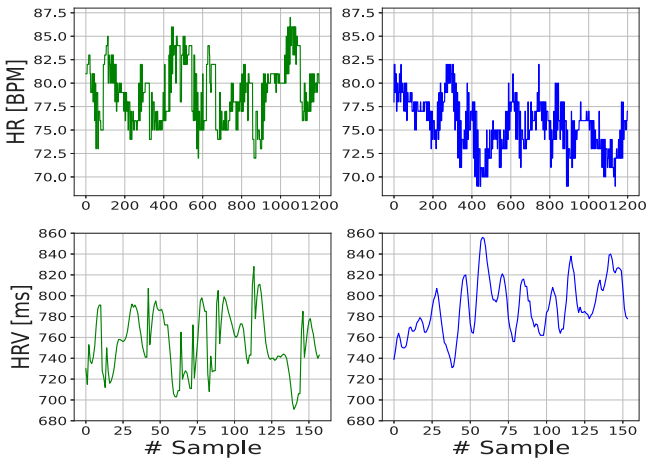


Fig. 6. Rest phase 0 (green) and rest phase 8 (blue) comparison of one subject during an experiment on the same day.

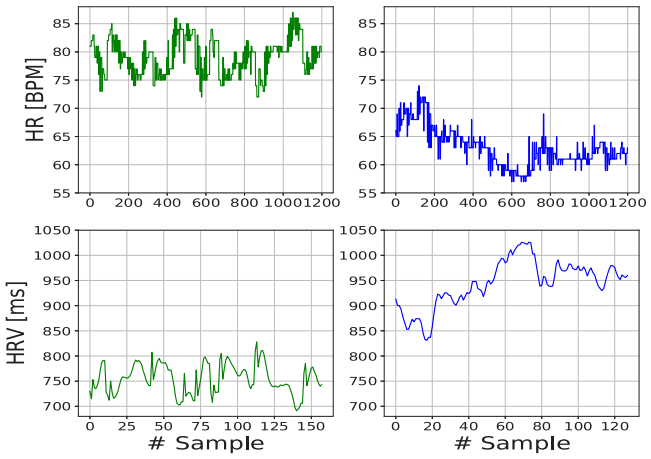


Fig. 7. Rest phase 0 comparison of one subject on different days: Day 1 (green) and Day 2 (blue).

day. This discrepancy is also observed in HRV, where the mean value is 754 ms on the first day and 944 ms on the second day.

Hence, one can conclude, that the HR and HRV vary significantly depending on the current state of the subject. Therefore no long-term reliable generalised baseline can be defined based on the acquired physiological stress response data from a subject.

Fig. 8 compares the rest phase data between different subjects. Hence, one can conclude that for further analysis towards physiological stress response differentiation and recognition, a normalisation of the data is necessary to compare the extracted features. Therefore, two normalisation techniques are applied: The first technique used is the maximum absolute value normalisation which scales all values to the maximum absolute value of the data. The second normalisation approach uses the global minimum X_{min} of the total recorded data as a factor $X_{normalised}$, given in the equation below.

$$X_{normalised} = 1 - \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

6. Data modelling

Two machine learning algorithms (Random Forest algorithm (RF) as described in Breiman (2001)), K-Nearest-Neighbours algorithm (KNN) as described in Mucherino, Papajorgji, and Pardalos (2009)), are investigated how precisely these algorithms can distinguish between rest

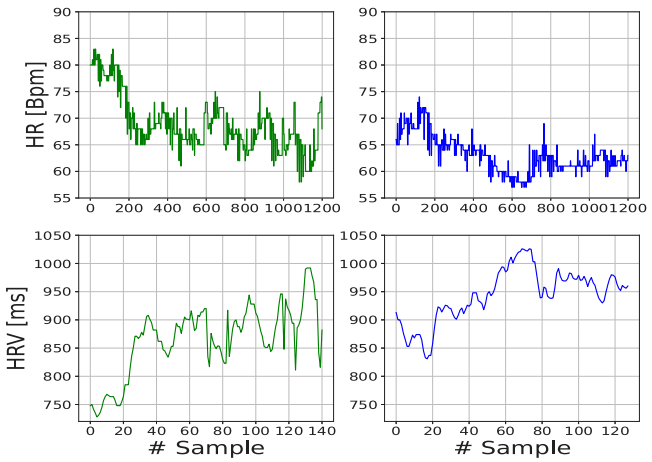


Fig. 8. Rest phase 0 comparison of two subjects during an experiment on the same day: subject 1 (green) and subject 2 (blue).

phase, easy task phase and difficult task phase based on a 67%–33% train-test split of the normalised data with two metrics: The measure of the ability to distinguish correctly between rest phase, easy task phase and difficult task phase is the accuracy. The robustness indicates the variance of the results of the algorithms by randomly changing the train-test split of the normalised data. For the training of the machine learning algorithms, the data of HR, HRV and RR are used, since the related research suggested that the RR is an indicator for physiological stress response in combination with other metrics (Goldberger et al., 2018; Reisman, 1997).

6.1. Task-rest differentiation

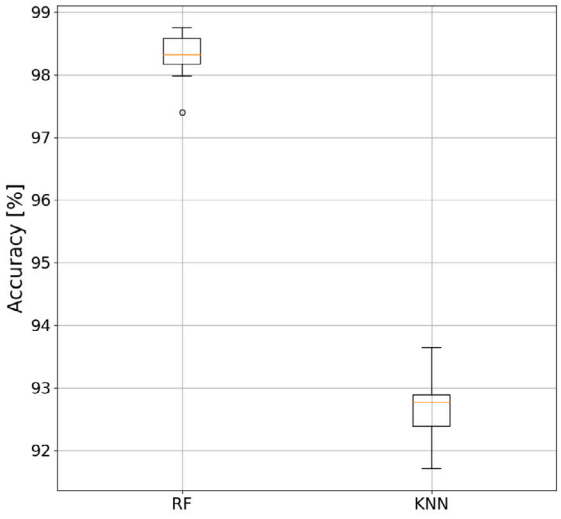
The first algorithm which is investigated is the KNN. The first training takes place with data from one subject during the experiment on the first day, to investigate the ability to differentiate between a task phase from a rest phase. The achieved accuracy is 90% of the easy task phase applying minimum normalised data whereas with maximum normalised data, the accuracy obtain 93%. Also, the RF is trained with the same data to investigate the ability to differentiate between task phase from a rest phase. The achieved accuracy is 95% of the easy task phase with the minimum normalisation and 98% applying maximum normalisation. Fig. 9 depicts the results of the investigation for the robustness of both investigated algorithms. Hence, one can conclude that the RF outperforms the KNN with the one subject train-test split and the RF performs better on maximum normalised data.

6.2. Stress level differentiation

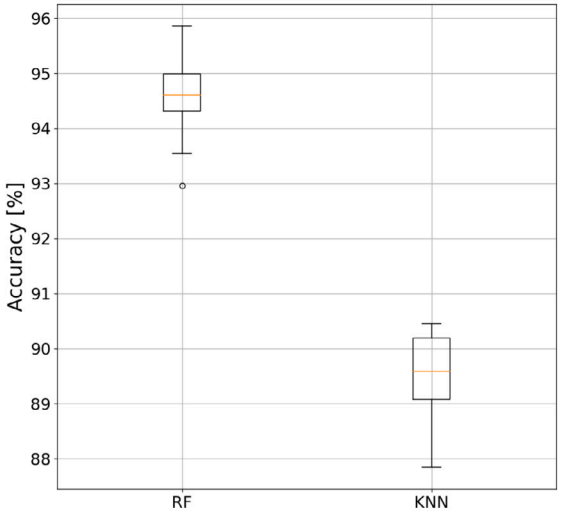
The KNN and the RF are investigated to distinguish between the easy task phase and the difficult task phase. First, normalised data from one subject is used and subsequently, normalised data of all subjects is used but from both experiments.

6.2.1. Normalised data from one subject

Both machine learning algorithms to be investigated are trained and tested on the data of one subject, as can be seen in Fig. 9. For KNN based on maximum normalised data the achieved average accuracy is 92.8% whereas for minimum normalisation data, the achieved average accuracy is 89.5%. For RF based on maximum normalised data the achieved average accuracy is 98.3% whereas for minimum normalisation data, the achieved average accuracy is 94.5%. One can conclude that the RF outperforms the KNN with both normalisation techniques on data from one subject.



(a) Robustness of algorithms for one subject train-test split for maximum normalisation for physiological stress response.



(b) Robustness of algorithms for one subject train-test split for minimum normalisation for physiological stress response.

Fig. 9. Robustness comparison of KNN and RF for one subject train-test split with different normalisation methods for stress detection.

6.2.2. Normalised data from all subjects

Both machine learning algorithms to be investigated are trained and tested on the data of all subjects, illustrated in Fig. 10. For KNN based on maximum normalised data the achieved average accuracy is 82.3% whereas for minimum normalisation data, the achieved accuracy is 74.7%. For RF based on maximum normalised data the achieved average accuracy is 98.3% whereas for minimum normalisation data, the achieved average accuracy is 94.8%.

Tables 2 and 3 depict precision, recall and F1-Score (as described in Goutte & Gaussier, 2005) to evaluate both investigated algorithms.

From the above results, one can conclude that both algorithms are capable of detecting physiological stress responses and differentiating

Table 2
Algorithm results for one subject data used for algorithms training and testing.

Algorithm results			
Labels	Precision	Recall	F1-Score
Rest RF	95%	97%	96%
Rest KNN	88%	88%	88%
Easy task RF	98%	95%	96%
Easy task KNN	87%	89%	88%
Difficult task RF	94%	93%	94%
Difficult task KNN	80%	75%	77%

Table 3
Algorithms results table for algorithms trained on all subjects but one and tested on an excluded subject's data.

Algorithm results			
Labels	Precision	Recall	F1-Score
Rest RF	84%	88%	86%
Rest KNN	81%	86%	83%
Easy task RF	42%	8%	14%
Easy task KNN	24%	7%	11%
Difficult task RF	49%	85%	62%
Difficult task KNN	51%	81%	63%

them between rest and task phases. Comparing both algorithms, the accuracy of the RF is in both cases better than the KNN.

7. Algorithm robustness

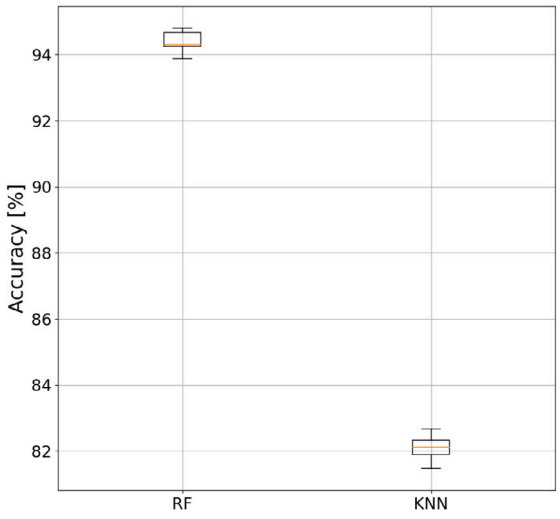
The boxplots illustrated in Figs. 9, 10 and 11 depict the K-fold cross-validation of both algorithms with different normalisation methods and data. The robustness of the algorithms is tested by repeatedly training the algorithms on different data (i.e. one subject train-test split, all subjects train-test split, ...) while changing the random train-test split at each iteration, to investigate the accuracy based on the provided data. The unbiased robustness results from Fig. 12 show an average accuracy for the RF algorithm of 56% and for the KNN algorithm an average accuracy of 43%. The robustness of the algorithms has also been tested for the biased case and the results can be seen in Fig. 10. The RF algorithm achieved an average accuracy of 94.3% with maximum normalised data and an average accuracy of 88.4% with minimum normalised data. The KNN algorithm achieved an average accuracy of 82.2% with maximum normalised data and an average accuracy of 75.2% with minimum normalised data.

The K-fold cross-validation is performed where the algorithms are trained on all data from the two experiments but one subject's data and tested with the excluded data.

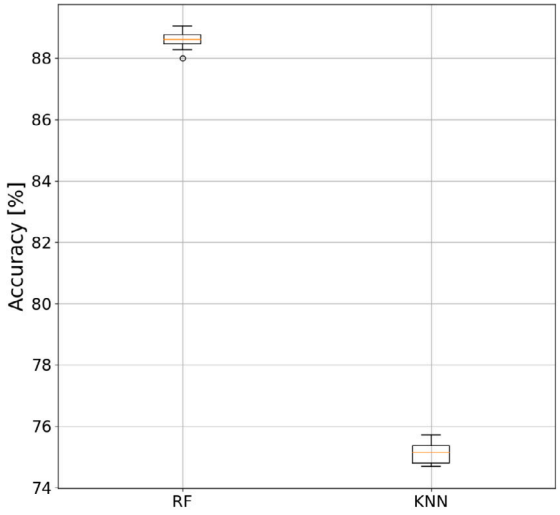
The KNN robustness evaluation achieved an average accuracy of 88%. The confusion matrix for KNN trained on all maximum normalised data but one and tested on the excluded data as well as the confusion matrix trained and tested on all data of the same subject is represented in Fig. 13.

The RF robustness evaluation achieved an average accuracy of 96%. The confusion matrix for RF trained on all maximum normalised data but one and tested on the excluded data as well as the confusion matrix trained and tested on all data of the same subject is represented in Fig. 14.

The evaluation of Figs. 10, 11, 12 indicates that the RF has overall higher accuracy and is more robust compared to the KNN in the experiments executed in this paper. Hence, the RF is more robust based on the experiments conducted in this paper and for physiological stress response detection and differentiation in the field of fine-motory assembly tasks.



(a) Robustness of algorithms for all subjects train-test split for maximum normalisation for physiological stress response detection and differentiation.

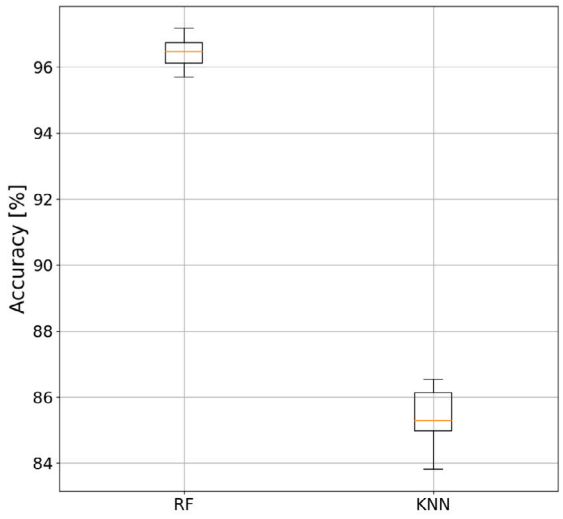


(b) Robustness of algorithms for all subjects train-test split for minimum normalisation for physiological stress response detection and differentiation.

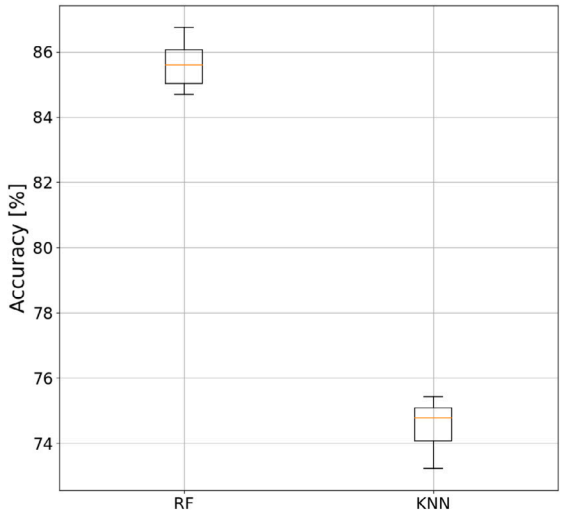
Fig. 10. Robustness comparison of KNN and RF for all subjects with different normalisation methods for physiological stress response detection and differentiation.

8. Results

In this paper, a commercially available wearable low-cost sensor is used to measure the physiological stress response (HR, HRV and RR) of subjects. Two machine learning algorithms, KNN and RF, are trained to automatically distinguish between the rest phase, easy task phase and difficult task phase based on normalised data. The KNN has a lower overall accuracy compared to the RF. The RF has a higher capability for both physiological stress response detection and differentiation. The highest accuracy achieved is between 75% and 90% with biased RF.



(a) Robustness of algorithms for one subject train-test split for maximum normalisation for the physiological stress response differentiation of task phase and rest phase.



(b) Robustness of algorithms for one subject train-test split for minimum normalisation for the physiological stress response differentiation of task phase and rest phase.

Fig. 11. Robustness comparison of KNN and RF for one subject train-test split with different normalisation methods for the physiological stress response differentiation of task phase and rest phase.

Table 4 shows all accuracies achieved with RF and KNN with different train-test data as well as normalisation methods.

The minimum normalisation reduces the accuracy of both algorithms significantly. One reason for this decrease in accuracy is based on the skew change and decrease of deviation of the distributions resulting from the normalisation. Applying maximum normalisation, the distribution of data has a higher similarity (standard deviation, skew, quantiles, etc.) to the distribution of the raw data compared to the minimum normalisation. The minimum normalisation reduces

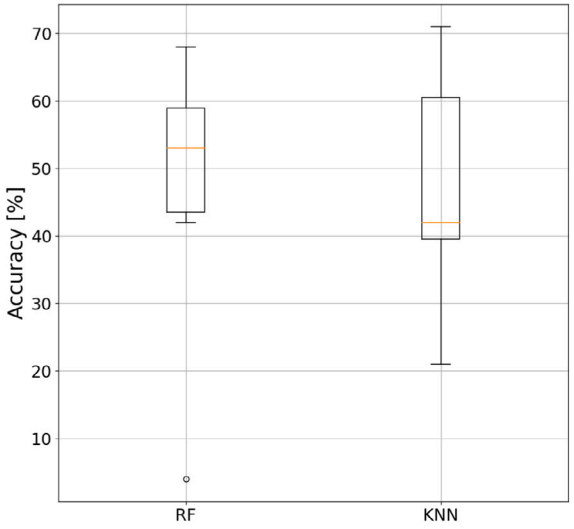
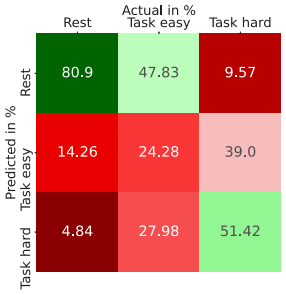
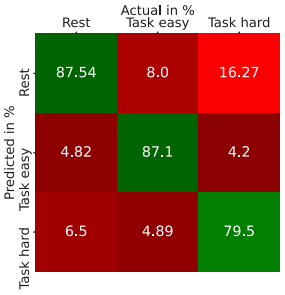


Fig. 12. Robustness comparison for algorithms trained on all subjects but one and tested on the excluded data with maximum normalisation.

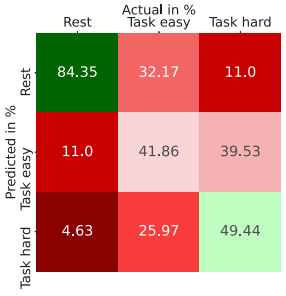


(a) Trained on all data except one set and tested on excluded set belonging to one subject.



(b) Trained on partial set of data of one subject and tested on excluded set of same subject.

Fig. 13. Confusion matrices for KNN.



(a) Trained on all data except one set and tested on excluded set belonging to one subject.

(b) Trained on partial set of data of one subject and tested on excluded set of same subject.

Fig. 14. Confusion matrices for RF.

the standard deviation of the distribution compared to the raw data distribution. This effect leads to worse robustness of both investigated machine learning algorithms.

Regarding the accuracy of the investigated algorithms compared to the subjective assessment of the subjects, Fig. 15 depicts the physiological stress response measured by HR (orange), HRV (blue), and RR

Table 4
Algorithm results for multiple test cases.

Accuracy of the investigated algorithms		
Test description	RF [%]	KNN [%]
One subject train-test split (maximum normalisation)	98	93
One subject train-test split (minimum normalisation)	93	93
All subjects data train-test split (maximum normalisation)	97	93
All subjects data train-test split (minimum normalisation)	89	75
Trained on all subjects but one; tested on excluded data (maximum normalisation)	95	84
Trained on all subjects but one; tested on excluded data (minimum normalisation)	85	74

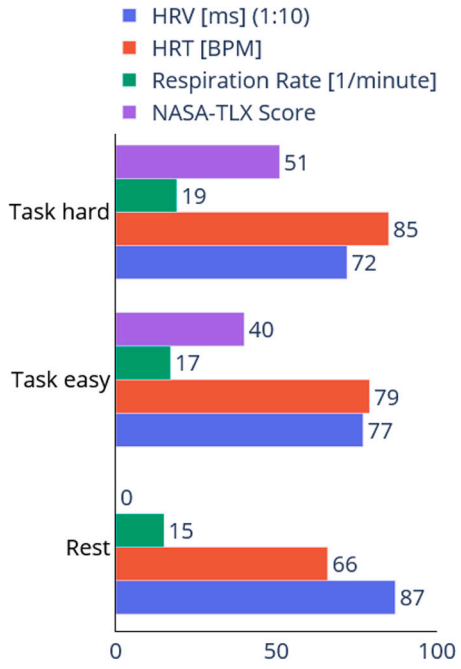


Fig. 15. Comparison of physiological stress response and subjective assessment method.

(green) with the subjective assessment method determined based on NASA-TLX score (violet).

From these results, one can conclude, that the detected physiological stress responses are compliant with the subjective assessment of the subjects.

9. Conclusion

This paper investigated whether it is possible to recognise and distinguish physiological stress responses caused by different workloads in assembly tasks using (i) commercially available wearable low-cost sensor that record the employee's heart rate, heart rate variability and respiration rate, and (ii) standard machine learning algorithms such as Random Forrest (RF) and K-Nearest-Neighbours (KNN)?

This research question is answered by the experiments conducted within this paper. In summary, it can be concluded, that the used commercially available wearable low-cost sensor acquires HR, HRV and RR data with sufficient precision as input for machine learning algorithms.

However, the data obtained must be normalised: Here, global maximum normalisation proved to be more suitable than global minimum normalisation.

If the subject's physiological stress response data is included in the training data for the machine learning algorithms, the accuracy of the RF is higher than the KNN. If this is not the case, the accuracies of RF and ANN are about the same.

In summary, it is possible to recognise and differentiate the physiological stress responses of employees via a commercially available wearable low-cost sensor and machine learning algorithms.

10. Outlook

In future, the authors of this paper plan to expand the target group of the studies to female subjects as well as subjects over 35 years of age and with different skill levels. Furthermore, the robustness of the two machine learning algorithms with regard to more complex assembly tasks through extensive hyperparameter tuning must be investigated. In addition, the performance of these two algorithms concerning an increased amount of data towards a universally applicable physiological stress response differentiation has to be researched. Other research topics, aside from investigating further algorithms, might include the challenge of interpreting HRV with the help of EEG data to interpret physiological stress response just from the HRV data.

Last, it is important to ensure that the use of this kind of personal data is not only in compliance with all existing data protection laws but can only be used to reduce the workload of individual employees and not to maximise profit.

CRediT authorship contribution statement

Markus Brillinger: Conceptualization, Data curation, Funding acquisition, Investigation, Validation, Writing – original draft, Writing – review & editing. **Samuel Manfredi:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Dominik Leder:** Data curation, Formal analysis, Investigation. **Martin Bloder:** Data curation, Formal analysis, Investigation, Methodology. **Markus Jäger:** Conceptualization, Formal analysis, Funding acquisition, Validation, Visualization. **Konrad Diwold:** Conceptualization, Formal analysis, Funding acquisition, Methodology. **Amer Kajmakovic:** Conceptualization, Formal analysis, Funding acquisition, Methodology. **Rudolf Pichler:** Funding acquisition, Writing – review & editing. **Martin Brunner:** Funding acquisition. **Stefan Mehr:** Funding acquisition. **Viktorijo Malisa:** Formal analysis, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Markus Brillinger reports financial support was provided by Pro2Future GmbH. Markus Brillinger reports a relationship with Pro2Future GmbH that includes: employment.

Data availability

Data will be made available on request.

Acknowledgements

This work has been supported by the FFG, Austria, Contract No. 881844: "Pro²Future is funded within the Austrian COMET Program Competence Centers for Excellent Technologies under the auspices of the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology, the Austrian Federal Ministry for Digital and Economic Affairs and of the Provinces of Upper Austria and Styria. COMET is managed by the Austrian Research Promotion Agency FFG".

References

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, *36*(7), 623–636. http://dx.doi.org/10.1016/j.ergon.2006.04.002.

Akmandor, A. O., & Jha, N. K. (2017). Keep the stress away with SoDA: Stress detection and alleviation system. *IEEE Transactions on Multi-Scale Computing Systems*, *3*(4), 269–282. http://dx.doi.org/10.1109/TMSCS.2017.2703613.

Alsuraykh, N. H., Wilson, M. L., Tennent, P., & Sharples, S. (2019). How stress and mental workload are connected. In *Proceedings of the 13th EAI international conference on pervasive computing technologies for healthcare* (pp. 371–376). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3329189.3329235.

Bachner, P. (2003). Clinical laboratory medicine. *Clinical Chemistry - CLIN CHEM*, *49*, 344–345. http://dx.doi.org/10.1373/49.2.344.

Báez, Y. A., Rodriguez, M. A., Limon, J., & Tlapa, D. A. (2014). Model of human reliability for manual workers in assembly lines. In *2014 IEEE international conference on industrial engineering and engineering management* (pp. 1448–1452). IEEE.

Bakker, J., Pechenizkiy, M., & Sidorova, N. (2011). What's your current stress level? Detection of stress patterns from gsr sensor data. In *2011 IEEE 11th international conference on data mining workshops* (pp. 573–580). http://dx.doi.org/10.1109/ICDMW.2011.178.

Baua (2020). Stressreport deutschland 2019: Psychische anforderungen, ressourcen und befinden. *Amtliche Mitteilungen der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin*, *4*, 16, arXiv:shorturl.at/fiyGH.

Boysen, N., Fliedner, M., & Scholl, A. (2008). Assembly line balancing: Which model to use when? *International Journal of Production Economics*, *111*(2), 509–528.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Breznik, M., Buchmeister, B., & Vujica Herzog, N. (2023). Assembly line optimization using MTM time standard and simulation modeling—A case study. *Applied Sciences*, *13*(10), 6265.

Calawa, R., & Smith, G. (2017). High volume automated spar assembly line (SAL). In *AeroTech congress & exhibition*. SAE International, http://dx.doi.org/10.4271/2017-01-2073.

Caroline Chanel, P. C., Wilson, M. D., & Scannella, S. (2019). Online ECG-based features for cognitive load assessment. In *2019 IEEE international conference on systems, man and cybernetics* (pp. 3710–3717). http://dx.doi.org/10.1109/SMC.2019.8914002.

Carrasco, G. A., & Van de Kar, L. D. (2003). Neuroendocrine pharmacology of stress. *European Journal of Pharmacology*, *463*(1–3), 235–272.

Chen, T., Yuen, P., Richardson, M., Liu, G., & She, Z. (2014). Detection of psychological stress using a hyperspectral imaging technique. *IEEE Transactions on Affective Computing*, *5*(4), 391–405. http://dx.doi.org/10.1109/TAFFC.2014.2362513.

Chi, Y. M., Jung, T.-P., & Cauwenberghs, G. (2010). Dry-contact and noncontact biopotential electrodes: methodological review. *IEEE Reviews in Biomedical Engineering*, *3*, 106–119.

Cho, I.-H., Kim, D. H., & Park, S. (2020). Electrochemical biosensors: Perspective on functional nanomaterials for on-site analysis. *Biomaterials Research*, *24*(1), 1–12. http://dx.doi.org/10.1186/s40824-019-0181-y.

Chung, M. K., Lee, I., & Yeo, Y. S. (2001). Physiological workload evaluation of screw driving tasks in automobile assembly jobs. *International Journal of Industrial Ergonomics*, *28*(3–4), 181–188.

Engström, J., Markkula, G., Victor, T., & Merat, N. (2017). Effects of cognitive load on driving performance: The cognitive control hypothesis. *Human Factors*, *59*(5), 734–764. http://dx.doi.org/10.1177/0018720817690639.

Fahr, A., & Hofer, M. (2013). Psychophysiologische messmethoden. *Handbuch Standardisierte Erhebungsverfahren in Der Kommunikationswissenschaft* (pp. 347–365). Wiesbaden: Springer Fachmedien Wiesbaden, http://dx.doi.org/10.1007/978-3-531-18776-1_19.

Finco, S., Battini, D., Delorme, X., Persona, A., & Sgarbossa, F. (2020). Workers' rest allowance and smoothing of the workload in assembly lines. *International Journal of Production Research*, *58*(4), 1255–1270.

Fridman, L., Reimer, B., Mehler, B., & Freeman, W. T. (2018). Cognitive load estimation in the wild. In *Proceedings of the 2018 Chi conference on human factors in computing systems* (pp. 1–9). http://dx.doi.org/10.1145/3173574.3174226.

Goldberger, A. L., Goldberger, Z. D., & Shvilkin, A. (2018). Chapter 2 - ECG basics: Waves, intervals, and segments. In A. L. Goldberger, Z. D. Goldberger, & A. Shvilkin (Eds.), *Goldberger's clinical electrocardiography* (9th ed.). (pp. 6–10). Elsevier, http://dx.doi.org/10.1016/B978-0-323-40169-2.00002-0.

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. *3408*, In *A probabilistic interpretation of precision, recall and F-score, with implication for evaluation* (pp. 345–359). http://dx.doi.org/10.1007/978-3-540-31865-1_25,

Hagmüller, M., Rank, E., & Kubin, G. (2006). *Evaluation of the human voice for indications of workload-induced stress in the aviation environment: Technical report*, (pp. 4–23).

He, C., Mahfouf, M., & Torres-Salomao, L. A. (2018). Facial temperature markers for mental stress assessment in human-machine interface (HMI) control system. In *ICINCO (2)* (pp. 31–38). http://dx.doi.org/10.5220/0006820700210028.

Johansen, K., Rao, S., & Ashourpour, M. (2021). The role of automation in complexities of high-mix in low-volume production – A literature review. *Procedia CIRP*, *104*, 1452–1457. http://dx.doi.org/10.1016/j.procir.2021.11.245, 54th CIRP CMS 2021 - Towards Digitalized Manufacturing 4.0. URL https://www.sciencedirect.com/science/article/pii/S2212827121011434.

Kalscheuer, F., Eschen, H., & Schüppstuhl, T. (2021). Towards semi automated pre-assembly for aircraft interior production. *Annals of Scientific Society for Assembly, Handling and Industrial Robotics*, 203–213.

Kern, C., & Refflinghaus, R. (2015). Assembly-specific database for predicting human reliability in assembly operations. *Total Quality Management & Business Excellence*, *26*(9–10), 1056–1070.

Kumar, N., & Kumar, J. (2016). Measurement of cognitive load in HCI systems using EEG power spectrum: an experimental study. *Procedia Computer Science*, *84*, 70–78. http://dx.doi.org/10.1016/j.procs.2016.04.068.

Kumar, N., & kumar, j. (2016). Measurement of efficiency of auditory vs visual communication in HMI: A cognitive load approach. In *Measurement of efficiency of auditory vs visual communication in HMI: A cognitive load approach* (pp. 1–8). http://dx.doi.org/10.1109/HMI.2016.7449168.

Kun, A. L., Heeman, P. A., Paek, T., Miller, W. T., III, Green, P. A., Tashev, I., et al. (2011). Cognitive load and in-vehicle human-machine interaction. In *Adj. Proc. AutomotiveUI 2011*. arXiv:http://clw.hciunh.org/.

Laring, J., Forsman, M., Kadefors, R., & Örtengren, R. (2002). MTM-based ergonomic workload analysis. *International Journal of Industrial Ergonomics*, *30*(3), 135–148. http://dx.doi.org/10.1016/S0169-8141(02)00091-4, URL https://www.sciencedirect.com/science/article/pii/S0169814102000914.

Lundberg, U., Granqvist, M., Hansson, T., Magnusson, M., & Wallin, L. (1989). Psychological and physiological stress responses during repetitive work at an assembly line. *Work & Stress*, *3*(2), 143–153.

Ma, L., Zhang, W., Chablat, D., Bennis, F., & Guillaume, F. (2009). Multi-objective optimisation method for posture prediction and analysis with consideration of fatigue effect and its application case. *Computers & Industrial Engineering*, *57*(4), 1235–1246.

MacDonald, W. (2003). The impact of job demands and workload on stress and fatigue. *Australian Psychologist*, *38*(2), 102–117.

Mahmad Khairai, K., Abdul Wahab, M. N., & Sutarto, A. P. (2022). Heart rate variability (HRV) as a physiological marker of stress among electronics assembly line workers. In *Human-centered technology for a better tomorrow: Proceedings of HUMENS 2021* (pp. 3–14). Springer.

Mehler, B., Reimer, B., Coughlin, J., & Dusek, J. (2009). The impact of incremental increases in cognitive workload on physiological arousal and performance in Young adult drivers. *Transportation Research Record: Journal of the Transportation Research Board*, *2138*, 6–12. http://dx.doi.org/10.3141/2138-02.

Melin, B., Lundberg, U., Söderlund, J., & Granqvist, M. (1999). Psychological and physiological stress reactions of male and female assembly workers: a comparison between two different forms of work organization. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *20*(1), 47–61.

Mucherino, A., Papajorgji, P. J., & Pardalos, P. M. (2009). k-Nearest neighbor classification. In *Data mining in agriculture* (pp. 83–106). New York, NY: Springer New York, http://dx.doi.org/10.1007/978-0-387-88615-2_4.

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. A. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian computer-human interaction conference* (pp. 420–423). http://dx.doi.org/10.1145/2414536.2414602.

Ramachandran, K. M., & Tsokos, C. P. (2015). Chapter 12 - Nonparametric tests. In K. M. Ramachandran, & C. P. Tsokos (Eds.), *Mathematical statistics with applications in R* (2nd Ed.). (pp. 589–637). Boston: Academic Press.

Ramasamy, S., & Balan, A. (2018). Wearable sensors for ECG measurement: a review. *Sensor Review*, *38*(4), 412–419. http://dx.doi.org/10.1108/SR-06-2017-0110.

Reisman, S. (1997). Measurement of physiological stress. In *Proceedings of the IEEE 23rd northeast bioengineering conference* (pp. 21–23). http://dx.doi.org/10.1109/NEBC.1997.594939.

Rey, D., & Neuhäuser, M. (2011). Wilcoxon-signed-rank test. In *International encyclopedia of statistical science* (pp. 1658–1659). Berlin, Heidelberg: Springer Berlin Heidelberg, http://dx.doi.org/10.1007/978-3-642-04898-2_616.

Rubio, S., Díaz, E., Martín, J., & Puente, J. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology*, *53*(1), 61–86. http://dx.doi.org/10.1111/j.1464-0597.2004.00161.x.

Samson, C., & Koh, A. (2020). Stress monitoring and recent advancements in wearable biosensors. *Frontiers in Bioengineering and Biotechnology*, *8*, 1037. http://dx.doi.org/10.3389/fbioe.2020.01037.

Saptari, A., Leau, J. X., & Mohamad, N. A. (2015). The effect of time pressure, working position, component bin position and gender on human error in manual assembly line. In *2015 International conference on industrial engineering and operations management* (pp. 1–6). IEEE.

Solange, A., Gordon, D., Ubel, F., Shannon, D., Berger, A., & Cohen, R. (1981). Power spectrum analysis of heart rate fluctuation: A quantitative probe of beat-to-beat cardiovascular control. *Science*, *213*, 200–222. http://dx.doi.org/10.1126/science.6166045.

Stanton, N., Salmon, P., Walker, G., Baber, C., & Jenkins, D. (2005). *Human factors methods: A practical guide for engineering and design* (1st ed.). United Kingdom: Ashgate Publishing Limited.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. http://dx.doi.org/10.1023/a:1022193728205.

Tempelmeier, H. (2003). Practical considerations in the optimization of flow production systems. *International Journal of Production Research*, *41*(1), 149–170.

Thorvald, P., Lindblom, J., & Andreasson, R. (2019). On the development of a method for cognitive load assessment in manufacturing. *Robotics and Computer-Integrated Manufacturing*, *59*, 252–266. http://dx.doi.org/10.1016/j.rcim.2019.04.012.