UNIVERSITY OF BIRMINGHAM

An adaptive human sensor framework for human–robot collaboration

Buerkle, Achim; Matharu, Harveen; Al-Yacoub, Ali; Lohse, Niels; Bamber, Thomas; Ferreira, Pedro

Document Version
Publisher's PDF, also known as Version of record

Link to publication on Research at Birmingham portal

An adaptive human sensor framework for human–robot collaboration

Achim Buerkle[1] · Harveen Matharu[1] · Ali Al-Yacoub[1] · Niels Lohse[1] · Thomas Bamber[1] · Pedro Ferreira[1]

Abstract

Manufacturing challenges are increasing the demands for more agile and dexterous means of production. At the same time, these systems aim to maintain or even increase productivity. The challenges risen from these developments can be tackled through human–robot collaboration (HRC). HRC requires effective task distribution according to each party's distinctive strengths, which is envisioned to generate synergetic effects. To enable a seamless collaboration, the human and robot require a mutual awareness, which is challenging, due to the human and robot "speaking" different languages as in analogue and digital. This challenge can be addressed by equipping the robot with a model of the human. Despite a range of models being available, data-driven models of the human are still at an early stage. For this purpose, this paper proposes an adaptive human sensor framework, which incorporates objective, subjective, and physiological metrics, as well as associated machine learning. Thus, it is envisioned to adapt to the uniqueness and dynamic nature of human behavior. To test the framework, a validation experiment was performed, including 18 participants, which aims to predict perceived workload during two scenarios, namely a manual and an HRC assembly task. Perceived workloads are described to have a substantial impact on a human operator's task performance. Throughout the experiment, physiological data from an electroencephalogram (EEG), an electrocardiogram (ECG), and respiration sensor was collected and interpreted. For subjective metrics, the standardized NASA Task Load Index was used. Objective metrics included task completion time and number of errors/assistance requests. Overall, the framework revealed a promising potential towards an adaptive behavior, which is ultimately envisioned to enable a more effective HRC.

Keywords Data-driven modelling · Human–robot collaboration · Machine learning · NASA-TLX · Wearable sensors

1 Introduction

The manufacturing landscape is moving towards the production of customized and personalized products [1]. This is mainly due to companies that prioritized their customer's needs, outperforming corporations that had largely focused on maximizing shareholder value [2]. High customization and personalization lead to an increase in uncertain production volumes, constant variant updates, and shorter production cycles [1]. Current automation technology, however, is often unable to meet the needs for flexibility and adaptability [1, 3]. This is due to the fact that complex assembly tasks require levels of reasoning, perception, and dexterity that

exceed the capabilities of conventional industrial robots [3]. As a result, an empirical study on the level of automation in the German manufacturing industry shows that approximately one third of the companies reduced their level of automation, as solutions were not flexible enough, and thus, not economical [4]. Alternatively, production domains with minimal deployment of automation technology exist. For example, most final assemblies in the automotive and aircraft industry are dominantly completed by manual operations [5]. However, due to the inability to maximize productivity of manual assembly systems, there is a strong motivation to increase the level of automation in these domains [6]. This is intended to overcome weaknesses associated with human workers such as being susceptible to high workload, fatigue, and stress [4].

An opportunity to bridge the gap between manual and fully automated systems can be human–robot collaboration (HRC) [3]. HRC combines the characteristic strengths of humans (perception, dexterity) and robots (precision,

✉ Achim Buerkle
a.buerkle@lboro.ac.uk

[1] Intelligent Automation Centre, Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, Loughborough, UK

fatigue-proof) to achieve common goals, which generates synergizing effects [3]. There are, however, several challenges that need to be addressed to successfully establish HRC. Since both human operators and robots are required to work in a shared workspace, safety systems need to ensure a worker's health and safety at all times [7]. Another challenge is the effective task distribution and allocation, based on state of a human/robot and the skills required [3, 8]. Thus, effective teamwork, as in the case of HRC, requires awareness of each member of the system [9]. This is essential for both establishing a safe environment and task planning/organization [10]. Ideally, the robot and the human would "understand" each other, which is particularly challenging due to the complementary differences [9]. These challenges can be tackled by equipping the robot with a model of the human to identify the current state, behavior, and intentions [9]. Hence, through the human model, a collaborative robot could estimate the human's goal/intention more accurately, and thus, adjust its behavior accordingly.

This paper presents a novel framework to establish data-driven models of the human for HRC. It combines physiological metrics (wearable sensors), subjective metrics (user interface), and objective metrics (time, quality) with machine learning to adapt to the uniqueness of each human operator and to integrate the gained insights into the collaborative system. Thus, it is envisioned to consider human factors, such as subject-specific characteristics and behaviors of individuals, which could ultimately lead to a higher system's performance.

2 Background

HRC is widely considered the highest level of human–robot interaction, since it includes jointly executed tasks [3, 6]. Moreover, it is viewed as the closest and most challenging method of interaction [3, 7]. The main goal of HRC is to combine the best of both worlds: joining the robot's precision, speed, and repeatability with the human intelligence, dexterity, and adaptability [3]. Thus, task scheduling and allocation is focused not only on dispatching tasks according to availability of a resource, but also on reaching an optimum based on different criteria such as assigning the most skilled entity, product requirements, and even energy consumption [11, 12]. In contrast to fully automated systems, the human worker introduces a new level of uncertainty and unpredictability [3]. This includes varying levels of worker's expertise, current state including health and fatigue, as well as comfort and ergonomic requirements [13]. Subsequently, these requirements are expected to lead to a constant and dynamic adjustment in task assignments, as well as during the execution of a task (adapted robot behavior) [5]. Overall, this enforces the need for accurate models of the human, which allow the collaborative robot to better understand its human partner.

In order to model human behavior, one needs to take a step back and establish a general understanding of what influences behavior. For this purpose, [14] developed the extended model of goal-directed behavior (EMGB). The key essence is that the majority of behaviors are functional and aim to achieve a certain goal. As shown in Fig. 1, Goals or Desires lead to intentions, which then lead to human behavior [14]. Moreover, the current behavior is influenced by past behavior (Frequency and Recency), as well as perceived behavior [15]. The Desire/Goal which leads to a behavior is described to be directly influenced by the person's Opportunity, Motivation, and Capability [14, 16], whereas Capability can be further split into Physical Capability and Psychological Capability [16].

In the context of HRC, human models have taken advantage of different aspects of this model. They can be grouped into different categories such as Marr's framework, which

Fig. 1 Extended goal-directed behavior (adapted) [14]–[16]
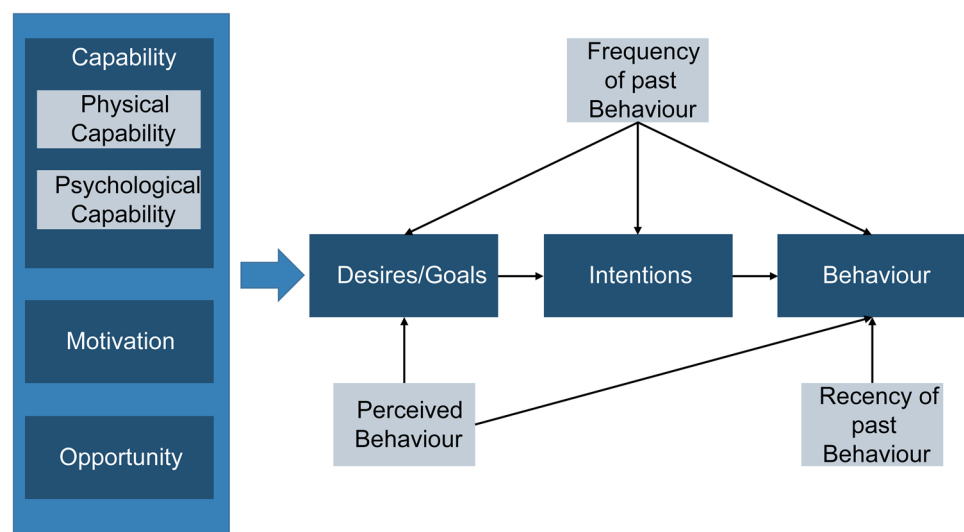
consists of three layers, namely computational, algorithmic, and implementation [17].

The main motivation for computational models (first layer) is to describe *what* the human is doing [17]. This is described to include clear and transparent mathematical functions [9]. For instance, probabilistic models are utilized, based on the goals of a task, to predict human intentions. These models achieved high accuracies for simple and intuitive tasks [18]. Yet, the accuracy is expected to decrease with an increasing complexity of tasks and activities. Moreover, HRC aims to take advantage of the human strengths such as adaptability and improvisation [3], which imply a higher level of complexity and thus potentially lower prediction results. Another type of computational models are knowledge-based models, which focus on the principle that current human behavior is based on previous behavior (ontological based models) [9, 19]. The assumptions for these models include that their domain knowledge of the human behavior is either complete or containing the most critical elements [9]. This, however, is widely considered unrealistic for tasks that exceed a certain level of complexity [20].

The second layer, containing algorithmic models, mainly focuses on the processes in human cognition or *how* the human is doing things [9]. This includes levels of reasoning, problem solving, and decision making [9, 17]. Despite various models being available, establishing algorithmic models remains challenging [9]. This is mainly due to the unobservable nature of the human thought process [9, 17]. Thus, models often remain in a fairly narrow context [9].

The third layer of Marr's framework consists of implementation or data-driven models [17], which aim to capture "real" physical or biological algorithms as they occur [9]. One of the applications is to capture limitations of human capabilities, such as how stress or the consumption of alcohol affects a person's performance [17]. Implementation models are described to be at an early stage in HRC [3, 9]. Yet, there are approaches, such as to apply electromyography (EMG) sensors to detect an operator's physical fatigue in HRC. In that case, a robot could assist to reduce the payload and the subsequent stress on human joints [21, 22]. This is intended to prevent injuries, as well as long-term health issues related to physical fatigue [21]. Moreover, it is taking advantage of the characteristic strengths of robots to cope well with high payloads, and thus helps to establish synergetic effects. Adding EMG sensor data has also shown promising potential to improve the mapping of force/torque and displacements in human–robot co-manipulation tasks [23].

In addition, brain-computer interfaces such as an electroencephalography (EEG) in HRC have gained a high research interest [3, 24]. In one application, an EEG was being used in conjunction with machine learning to decode human movement intentions in the brain, prior to their execution [25]. Thus, the collaborative robot could adjust its behavior accordingly.

Despite these approaches, there is still a large untapped potential for data-driven models of the human in HRC. Apart from measuring the human physical state with an EMG, it would be also beneficial to assess the human mental state in HRC, which directly influences the human goals and behavior. The Health and Safety Executive 2018 [26] suggests that high levels of mental workload can occur during complicated and therefore demanding tasks, as well as during repetitive, monotonous, and frustrating tasks [26]. This can lead to fatigue, stress, and poor job satisfaction [26]. Stress and fatigue are often related to a decrease in task performance [27]. Subsequently, there is an increased interest in human factors research such as measuring the perceived workload through physiological metrics. In general, physiological metrics offer a variety of measurements on how the human body responds to stimulations such as work tasks [13]. Although physiological data can provide insights on emotional states, due to the peculiarity of human beings, interpretations are often challenging [28]. Therefore, it is often required to be interpreted in a context or alongside with other metrics or an artificial intelligence (AI) [25, 28]. Moreover, more than one sensor technology can be used to measure a metric [13, 28]. An overview of potential physiological sensors is presented in Fig. 2.

Brain activity can be measured with the aforementioned EEG. This technology requires electrodes to be placed across the scalp to measure activities in the cerebral cortex [29]. Overall, five main frequency bands are distinguished. The higher the frequencies, the more active states are associated with it: from being drowsy, towards relaxed, then active thinking and focus, to alertness [25]. Lowered alpha waves (8–12 Hz) and increased gamma waves (30–100 Hz) are associated with higher workloads and stress [30, 31].

Pupillometry includes the analysis of pupil diameter, blink rate, and constant motion, known as saccades [32]. An increased period between blinks, for example, can be linked to a higher mental workload [13]. Eye-tracking is often combined with other physiological sensors [32].

Nose temperature: the human internal temperature correlates with physical and psychological states [33]. The reaction to stimuli increases or decreases blood flow which leads to variation in skin temperature. The variability in skin temperature is often measured with thermal imaging [34]. Among other facial regions, the nose tip is regarded to provide the most consistent indications of stress. Once stressful conditions apply, the nasal temperature decreases [33].

As described before, an EMG detects the contradiction of muscles. When placed in the face, it can detect tensions occurring from a clenched jaw. These readings can be associated with stress and a higher workload [28].

Fig. 2 Human physiological interfaces (adapted) [28]
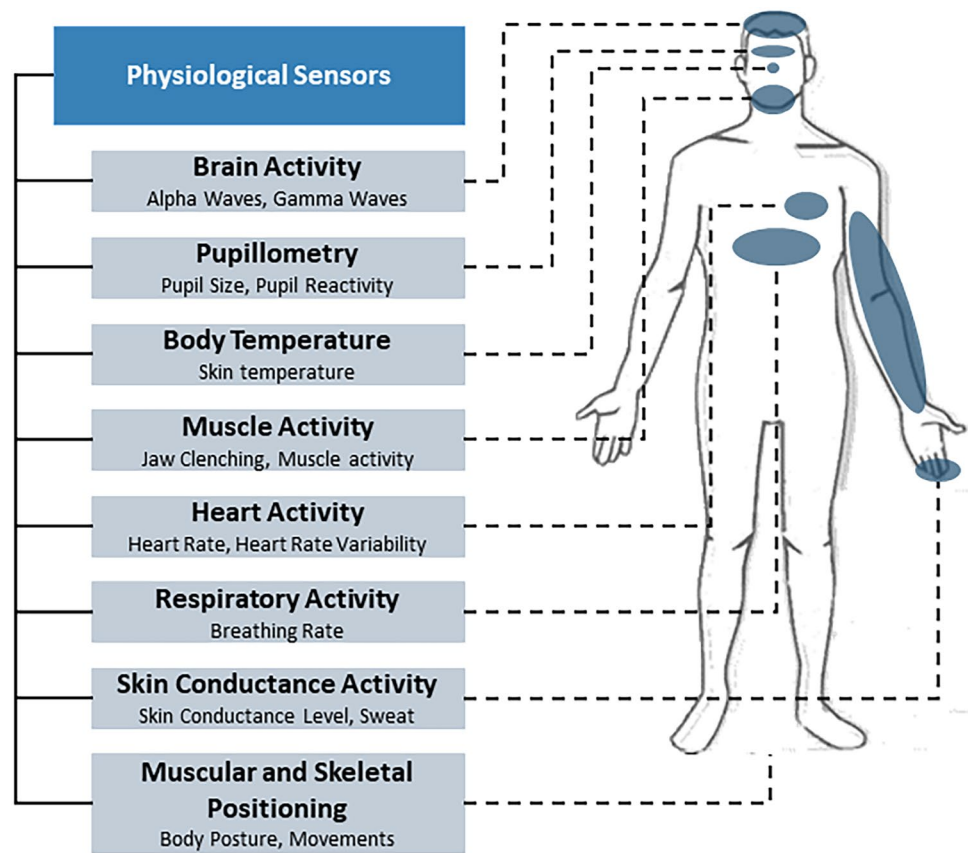
Cardiovascular signals are measured and monitored with an electrocardiogram (ECG). The ECG can detect a heart's responses to stimuli. Responses can be an increased/lowered heart rate, heart rate variability, blood pressure, and blood volume pressure [28]. The recorded data can be interpreted as the occurrence of stress, to measure mental effort, and various emotional sates. Generally, an increase in heart rate over time or a decrease in heart rate variability is linked to a higher mental workload [28, 29]

Respiratory activity such as breathing rate and breathing rate variability is strongly linked to cardiovascular activity. Measurements track the expansion and contraction of the chest [28]. A decrease in breathing rate in combination with a higher intensity is attributed to higher levels of mental workload [35].

Skin conductance activity (SCA) or galvanic skin response relies on the electrical conductivity that occurs when the human body produces sweat [28]. Larger amounts of sweat lead to a higher conductivity. Thus, an increase of SCA is associated with an increase in mental workload [35, 36]. A disadvantage, however, is that SCA is typically measured on the finger tips [28]. This could conflict with a worker's ability to perform a task in a human–robot collaborative scenario.

Muscular and skeletal positioning tracking can be used as a physiological measurement to link a person's posture to a mental state [28]. Yet, the linkage between specific postures and mental workload is currently considered challenging [37].

Although physiological sensor data is often grouped in categories, which might lead to the impression that consistent interpretations might be established, they often lack a monolithic interpretation, due to a wide variety of subject-specific characteristics and noise within the data [28, 29, 38]. Nevertheless, with advances regarding machine learning, there is the opportunity to further utilize these sensor technologies, and to establish a more accurate, real-time model of the human.

3 Framework architecture

In this section, an overview of the proposed adaptive human sensor framework for human–robot collaboration is presented. Since HRC is often applied in a manufacturing context, it is essential to not solely focus on the human factors, but to consider the collaborative robot's configuration and product requirements as well, to obtain an optimal
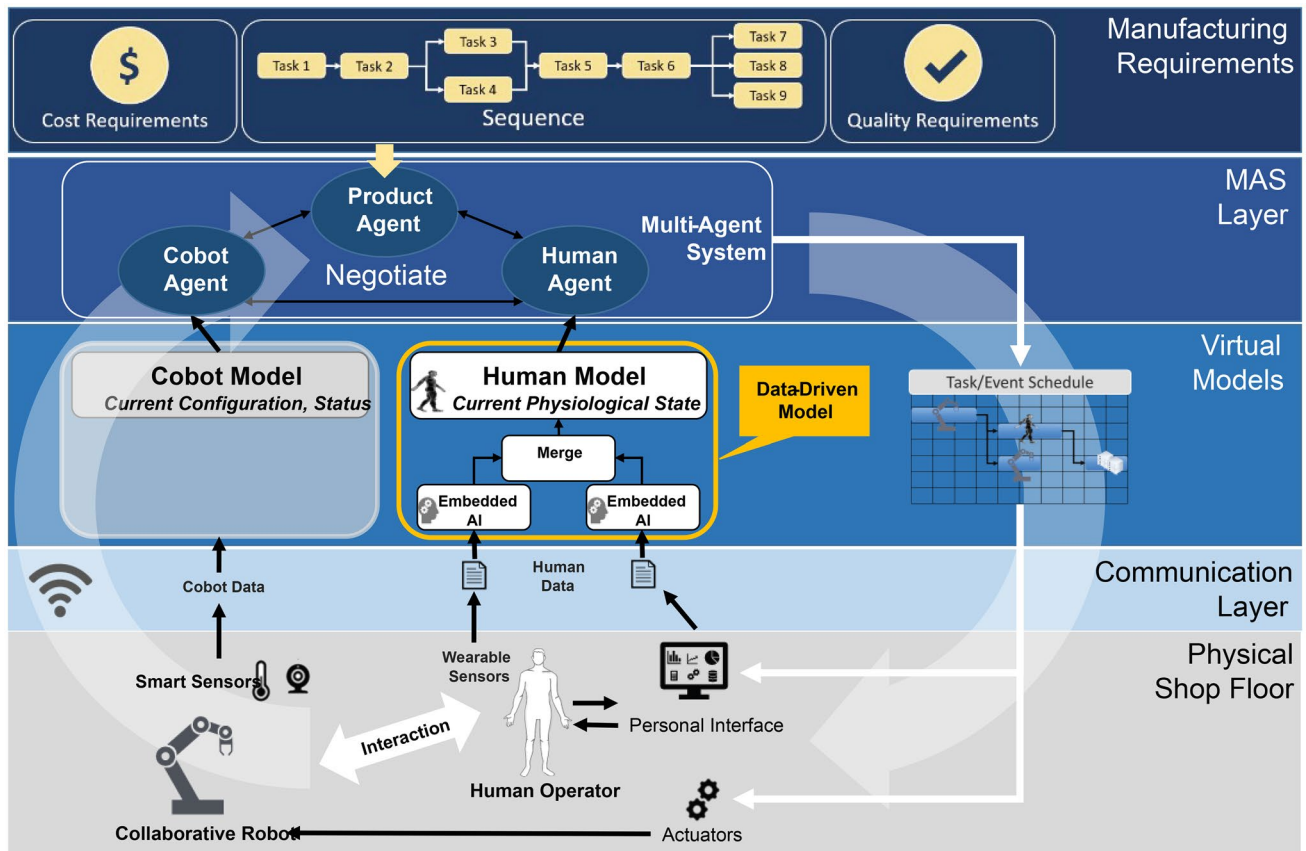
Fig. 3 Top level architecture of the framework

solution. Thus, awareness of all members in the collaborative system is required. On a top level, this could be organized by existing technologies, such as agents, as shown in Fig. 3. Essentially, agents form a digital representation of each physical entity to model the state, goals, and behavior in the digital domain [39].

While the concept of using agents in HRC itself is not new [3, 40], the stored information and behavior could be modelled more accurately, based on data-driven models, especially of the human operator. This would offer to model the characteristic strengths and weaknesses of humans, such as, for example, the dynamic nature of perceived workloads. These perceived workloads can occur either during very demanding tasks, or on the opposite scale, during monotonous and repetitive tasks. Overall, this is envisioned to increase productivity of the setup while maintaining the system's overall flexibility.

The product agent initiates the global optimization and task distribution. Its main goal is to be manufactured to the quality standards of the company. The quality requirements can only be achieved if the required skills are matched to the product requirements. Consequently, the product would aim to allocate the most skilled agent

(human or robot). This, however, could conflict with the required assembly time and subsequent costs. Therefore, the conflict between achieving a high quality and meeting the set order deadline needs to be implemented. Thus, overall, the product agent represents the manufacturing requirements or objective metrics. Objective metrics in human-centered experiments are commonly recorded as quantitative measures to establish comparable values across different individuals [41]. They are often used for the rating of task performance as measurements for completion time, accuracy, gauging error, and successfulness of the task [41].

The collaborative robot (cobot) agent holds the skills and goals of the robot's current configuration. This configuration does not only include software and pre-programmed task sequences. It is also based on the physical configuration of the robot, such as the tool attached to the robot's end-effector. Changing the configuration results in manual efforts to re-equip and reprogram a robot. Hence, the cobot-agent's goal is not only to perform mainly standardized tasks (e.g., pick and place), but also to minimize reconfiguration as much as possible, which is aligned with robot's characteristic strengths to cope well with repetitive tasks. The

opposite, as in adapting to changes, is widely considered a robot's weakness. Therefore, the cobot-agent aims to avoid reacting to unexpected events. This duty would be shifted over to the human-agent.

The human-agent is considered the most complex agent in the system, since human operators have different individual behaviors, skills (qualification, experience), and goals. Moreover, these behaviors and skills are expected to change over time. Consequently, an agent would be required to continuously learn the characteristics for each individual human-being. It is also required to adapt to new human behaviors. In order to measure human states (physical and psychological capability) that are related to goals and behaviors, two types of metrics need to be acquired: subjective measurements and physiological data [28]. Individually, the significance/impact of each metric is rather limited due to the complexity of the human [28]. Advantages and disadvantages of each metric are analyzed in the following.

Subjective measurements typically include self-report measures such as questionnaires, surveys, and scales [36]. In the current context of predicting perceived workload, the NASA Task Load Index (TLX) was chosen, as it incorporates subjective ratings among six domains of perceived workload. This includes physical demand, mental demand, temporal demand, effort, performance, and frustration [42]. A disadvantage of any subjective metric, however, is that data is collected in retrospect, in which participants are expected to remember, interpret, and explain experiences, which is susceptible to cognitive biases [43]. Yet, standardized subjective metrics such as the NASA-TLX aim to eliminate this bias to the utmost [42].

Physiological measurements can be conducted with a variety of sensors, introduced in Sect. 2. The sensors each offer individual advantages and disadvantages, depending on the context. However, the sensory data often lacks a single monolithic interpretation [28]. In order to cope with this challenge, machine learning algorithms are utilized to process and interpret physiological measurements, while minimizing manual programming and fine-tuning efforts [21].

Figure 4 provides an overview of the proposed human-agent incorporating a data-driven model, which continuously acquires and adapts to the human operator. The model aims to provide standardized interfaces, plug-and-play of new sensor technologies, and self-learning/adjusting capabilities, based on the acquired data.

3.1 Physical layer

At the lowest level, in the physical domain, data is acquired from physiological sensors. In the following experiment, a mobile electroencephalogram (EEG), an electrocardiogram (ECG), and a respiration sensor are included to predict perceived workloads. This sensor selection is based on the fact
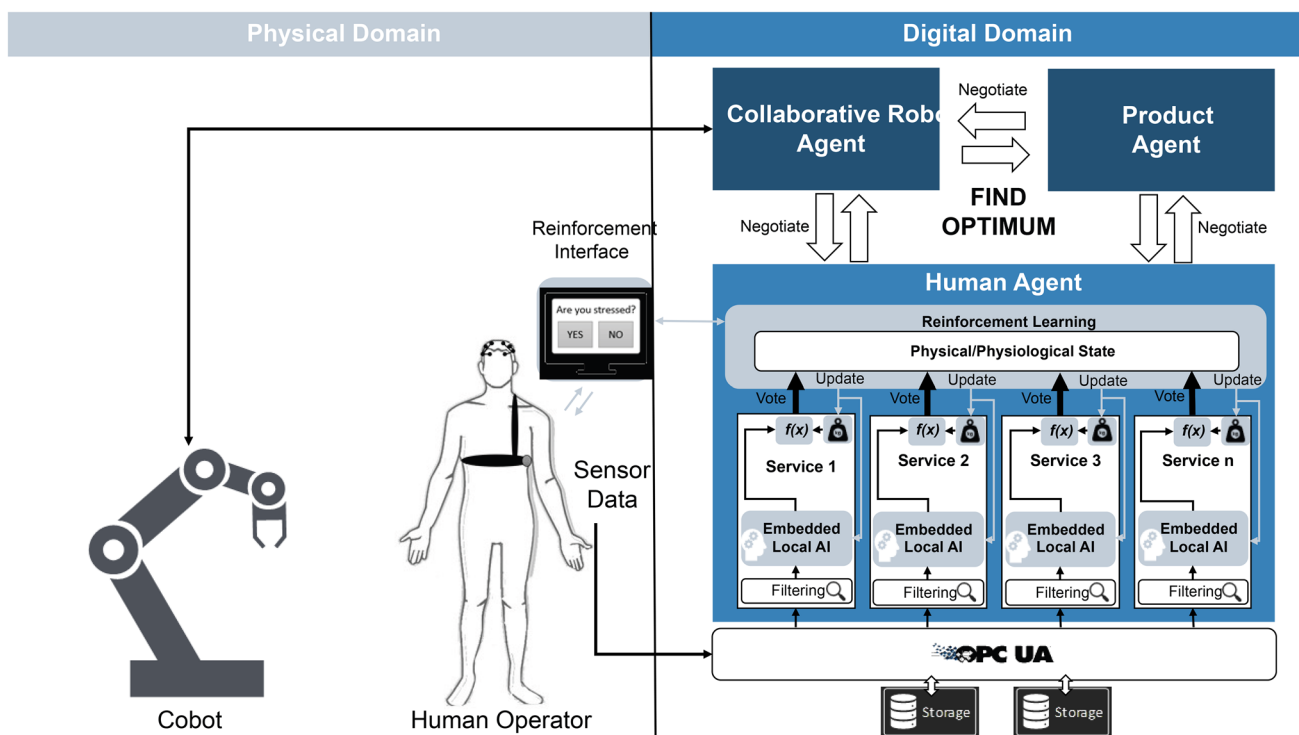

Fig. 4 Human-agent sensory framework

that they delivered promising results in previous research [13, 25, 28]. However, the selection is not limited to these sensors; other physiological sensors can be used, depending on their suitability for the context.

Generally, the main goal should be to minimize the intrusiveness of sensors and their subsequent interference with an operator's ability to perform tasks. Ideally, the sensors would allow a hands-free measurement, as this allows for operators to complete their tasks without hindrance.

3.2 Communication layer

In a second step, the sensor data is streamed through the communication layer. In the current setup, the OPC Unified Architecture (UA) is chosen, as it as a platform independent service-oriented architecture [44]. Thus, this allows for adding or removing sensors to the framework, when needed (plug-and-play). Moreover, OPC UA enables communication security, information modelling tools, and the ability to define a so-called value transmission. Within the value transmission, data structures and sampling intervals can be defined. This is particularly useful for processing human physiological data, since it often follows sinusoidal patterns, which requires an interval-based wave transform, before it can be interpreted [28]. In addition, the recorded data can be stored in databases, which allows for collecting data during a manufacturing process and to train machine learning classifiers both offline and online. An important consideration when storing physiological data needs to be that it is classified as sensitive personal data [28]. Therefore, it is essential to store the data in compliance with data protection regulations.

Overall, the selection of an appropriate technology to implement this communication layer is one of the crucial tasks regarding the complexity, speed, and performance of industrial distributed systems [45].

3.3 Data analysis and interpretation layer

The data analysis layer utilizes machine learning to extract relevant patterns and to predict a human state, which influences the human-agent's behavior, as depicted in Fig. 4. Given that each physiological sensor requires one or more data processing services, the data-driven model needs to allow for multiple services being plugged-in, as well as merging their results. The processing is executed in three steps: filtering, an embedded AI, and a weighted fuzzy logic-based voting at the end.

Depending on the sensor technology, the data preparation/filtering might be more intensive. As an example, EEGs are considered delicate devices that tend to measure large quantities of external noise [31]. Therefore, high filtering efforts are expected. ECG devices, on the other hand, often already provide absolute values of the heart rate and heart rate variability, which require little to no filtering [28].

After the filtering and preparation, the data can be interpreted. For this purpose, three main processing and machine learning techniques were identified, which can be applied based on the data complexity and volume, shown in Fig. 5.

3.3.1 Standard methodology

A common methodology for processing physiological data includes signal acquisition with a sensor, before splitting the data into intervals. Based on the sensor sampling rate and the context, different window sizes can be chosen. EEGs typically have a high sampling rate, which would allow for using smaller detection windows [28], whereas ECG sensors would need a larger window, since a heart rate of < 60 beats per minute implies less than a heartbeat per second. If an application is time critical, the window size needs to be minimized, as it adds to the processing time required.

Fig. 5 Human sensory data processing techniques
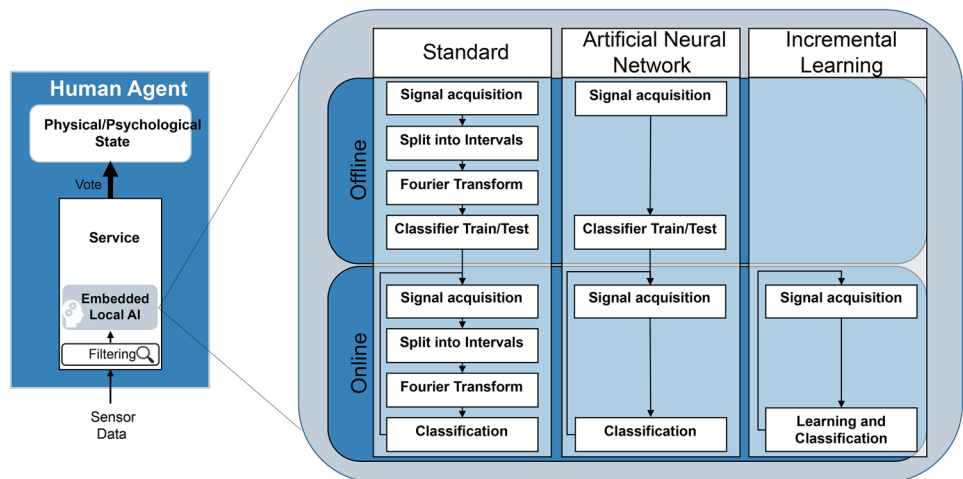
Afterwards, some physiological sensor data requires processing via a wavelet transform, often Fourier transform. This is mainly due to the fact that many physiological sensors provide data in sinusoidal patterns [28]. Consequently, the absolute values for the underlying frequencies can be extracted. These values are used to train and test a machine learning classifier offline, before its use in an online application. In literature, it is described that the processing technique such as fast Fourier transform can have a greater impact on the detection accuracy, than the classifier itself [46]. Nevertheless, the selection of a suitable classifier is essential to optimize processing time and to avoid overfitting. Many applications based on this methodology use Support Vector Machines (mostly with a Gaussian kernel) [28]. The frequent application of this classifier suggests it is somewhat considered state of the art. After training and testing of the classifier, it can be applied online.

The main advantage of this methodology is the high transparency, as the Fourier transform results can be visualized and manually inspected. Disadvantages, however, are the high manual tweaking and programming efforts. Moreover, one of the main issues identified in [46] regarding the classification of physiological data is the deviation between the classification accuracy offline and an online application. This can even be observed during different recording sessions for the same individual [46].

3.3.2 Artificial neural network

In contrast to the common methodology, there are also approaches in which raw sensor data is directly streamed into a classifier [25]. Thus, the classifier needs to cope with the noise and learn to identify the relevant constellation of features by itself [46]. In order to handle the complexity and even the time-based shape of the data, classifiers need to be fairly advanced. Mostly neural networks are utilized for these purposes, including specified subcategories of neural networks such as recurrent neural networks for time series-based prediction [25].

One of the main advantages of this approach is the low fine-tuning effort. The framework is coping with the raw data and adapting to the noise, it might even identify patterns that could not be recognized in a manual process. Disadvantages of this methodology can be the large amount of training data required to train a neural network. Moreover, after the training process, the model itself is considered a "Black Box," meaning the classification logic cannot be visualized and thus it is often difficult to comprehend what constellation led to a decision. A trained neural network is less likely to have (severely) varying detection accuracies between offline and online sessions, and still, the detection accuracy can decrease during new recording sessions [46].

3.3.3 Incremental learning

Incremental learning algorithms aim to address the limitations of varying classification results between offline and online systems by performing learning and classification operations online only. Moreover, it offers the ability of life-long learning, which allows to further tune the model's structure and performance over time. Thus, it allows to adapt to subject-specific characteristics within human sensory data [22, 47]. In addition, no prior knowledge about the number of classes or instances is required, which allows to drastically lower the manual programming efforts [47]. There are, however, disadvantages to incremental learning such as the plasticity-stability dilemma, which implies the model has to obtain new knowledge, while at the same time, it must not forget previous knowledge. Moreover, the more complex a model becomes, the longer it will take to perform the learning and classification operations [22, 47]. Subsequently, these issues need to be taken into account, when applying incremental learning for human sensory data. Nevertheless, incremental learning offers the potential to significantly reduce manual programming and fine-tuning efforts, when adding and integrating various physiological sensors to the human-agent.

3.4 Reinforcement learning layer

After processing available human-agent sensory data using suitable data processing services, the sensor data can be mapped with relevant subjective analysis through the reinforcement interface. This also enables to have a different set of machine learning models (services) that use the same sensor data and features to predict different subjective metrics. Consequently, each service can be weighted based on its prediction performance. Such an approach is called ensemble learning, which is a set of weak learners, where a prediction is based on voting. The ensemble learning establishes a weighted voting of each sensor/service, and then the final prediction is made based on the majority of votes. The weight of a service ranges from 0 to 100. A correct prediction of the human current state will result in a minor increase in weight. Yet, an incorrect prediction will lead to a more drastic decrease in weight (penalty). To determine the correctness of the predictions, the classification results of each sensor are compared with subjective metrics, which are collected through a user interface. In the current context, it is displaying a digital version of the NASA-TLX (perceived workload). The overall subjective score is then used as input for this reinforcement learning (adjusting the weights).

When establishing a subjective interface, it is essential to minimize the frequency of subjective metric requests, as they would distract a human operator from their current task.

Also, a constant demand for subjective feedback could lead to an increase in stress and thus, higher perceived workloads. Therefore, a subjective metric user interface could only be triggered, once the individual sensor weights show a large deviation.

3.5 Human physiological state layer

In the proposed framework, the state within the human-agent is represented by a set of services that adapt to the human behavior through the weighting mechanisms. The human-agent will then change/adjust its behavior and negotiations with other agents, accordingly, based on the physiological state. In the current context, a high perceived workload will result in the human-agent negotiating a change in tasks for its human operator or to trigger an assistant request of the robot, which aims to reduce the perceived workload of the operator. A low perceived workload will lead to the human-agent requesting more tasks or a more demanding task for its human counterpart. Thus, the human-agent aims to keep the perceived workload at an optimal level.

Since the product agent stores information on quality and assembly time, it can provide insights on (objective) task performance. The task performance could then be correlated with the physiological state of the human operator. Moreover, the physiological state of a human operator could also be linked with the weights of each physiological sensor. Thus, the sensor weights would create an easily interpretable fit-for-purpose sensor overview. This self-learning capability of the system could offer the potential to reduce manual programming and tweaking efforts. To test and validate the proposed adaptive framework, an experiment was conducted including two assembly experiments.

4 Experimental design and setup

The experimental design aims to demonstrate the framework's adaptive behavior by creating a comparison between two distinguished scenarios. In the first scenario, participants are expected to perform a manual assembly task without assistance of the robot. Higher perceived workloads are expected, as the task must be completed as quickly as possible, and participants are to remember all assembly steps. The second scenario contains a robot-assisted assembly task. Pick-and-place tasks are completed by the robot, which delivers the individual parts in the correct order. The human operator is tasked to assemble the pieces correctly. This highly assisted task is expected to create lower perceived workloads for participants. Thus, this would allow for the framework to predict different perceived workloads during the two scenarios. To avoid bias, participants with an odd number would perform the manual assembly task first and then the collaborative task. Participants with an even number, on the other hand, would perform the collaborative task first, and then the manual task.

Throughout the experiment, participants were equipped with wearable sensors, namely an Emotiv Epoc + EEG and a BioHarness 3 (ECG and Respiration Sensor). Figure 6 shows the task sequence, as well as the experimental setup in a shared human–robot cell, composed of a UR10 collaborative robot with a set of parts to be assembled. To correlate objective, subjective, and physiological metrics, all three are needed to be incorporated into the experiment. Therefore, objective measurements such as task performance are acquired by the experimental supervisor. Subjective metrics are gathered in the form of the NASA-TLX questionnaire after the completion of each scenario. Physiological data is collected in the form of brainwave activity (EEG), heart rate (ECG), and breath rate. The sensory data was automatically collected and stored through the proposed framework. Afterwards, machine learning was applied to identify features within the physiological data that correlate with the perceived workloads.

5 Results and discussion

This section presents the results of the validation experiment, as well as the learning outcomes of the proposed framework. In total, 18 participants took part in the experiment, 10 female and 8 male participants between 22 and 59 years of age (average: 25.5 years). Almost half of the participants had no prior experience working with industrial robots. Lack of experience has been indicated to affect a participant's performance in experiments [41]. Yet, no detectable correlations between lack of experience and performance could be found in the experimental data. Throughout the experiment, data from three different sensors was collected: namely EEG, ECG, and respiration sensors.

For the analysis of the EEG data, the channels F7, F3, FC5, AF3, AF4, FC6, F4, and F8 were used, since literature suggests higher detection accuracy, when measuring workload-related signals over the frontal lobe [31].The analysis results of the processed EEG data, using a notch filter (to filter power line noise) and fast Fourier transform, showed a difference in alpha waves (8–12 Hz) during both scenarios across all participants, in which lower alpha waves occurred during the manual task and higher alpha waves during the collaborative task, as shown in Fig. 7 part A. This was expected, as literature suggests lower alpha wave occurrences during an increased mental workload [30]. Similarly, higher amounts of low-range gamma waves (> 30 Hz) occurred during the manual task, and lesser occurrences were detected during the collaborative task, shown in Fig. 7
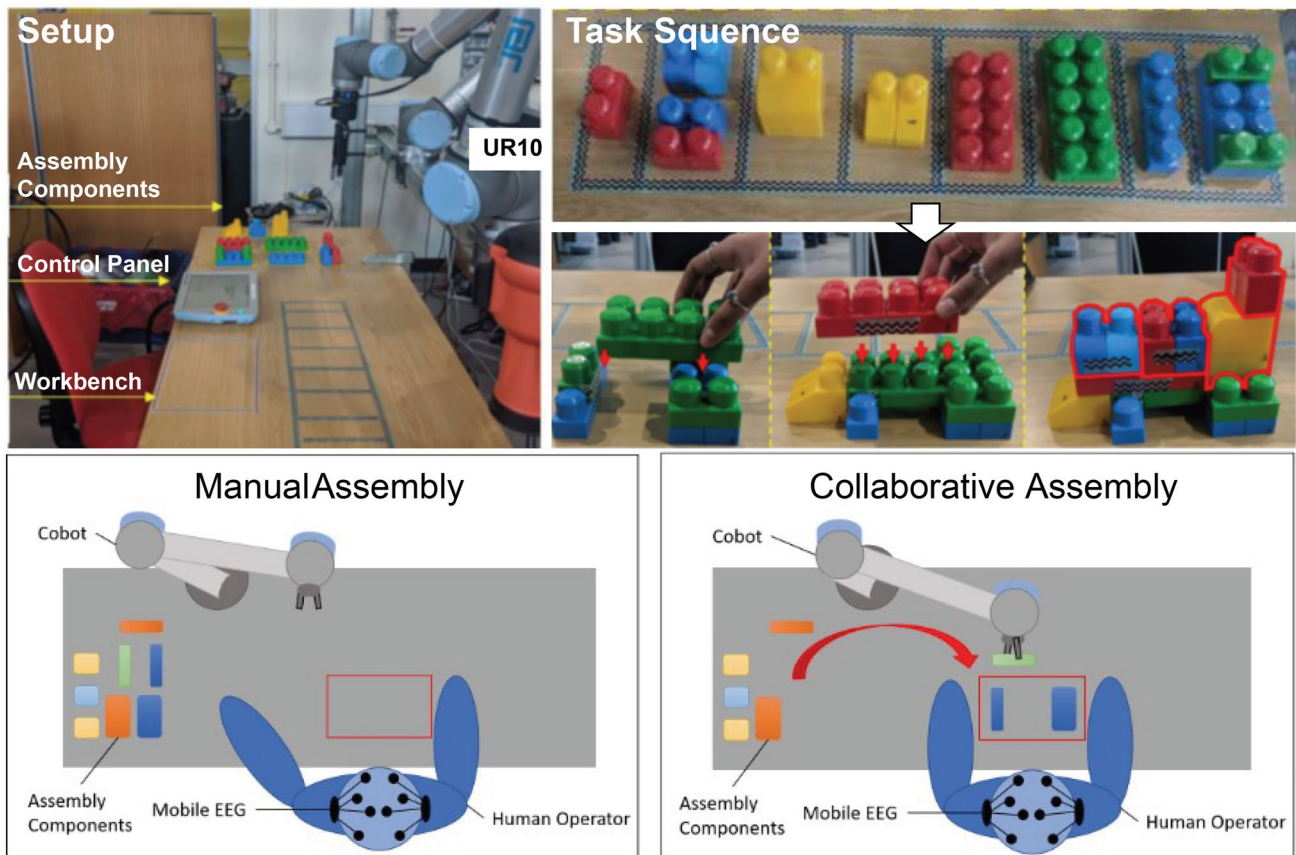
Fig. 6 Experimental design and setup

part B. This result was also anticipated, as an increase in gamma waves can be observed during an increased mental/perceived workload [31].

The ECG data showed a common pattern among different participants of slightly higher heart rates during the manual task and somewhat lower heart rates during the collaborative task, as shown in Fig. 7 part D. The only exception is participant 17 who showed an increased heart rate during the collaborative task. Overall, these results are expected, since literature suggests an increase in heart rate during a higher perceived workload [13, 28]. Moreover, a difference between participant's individual heart rates was also anticipated. For participant 6, heart rates between 48 and 57 beats per minute were observed. Opposing to that, participant 10 showed a heart rate of 108 to 117 beats per minute. Thus, this highlights the uniqueness of human beings and the subsequent need to establish individual models in order to correctly interpret these values.

In contrast to EEG and ECG data, the respiratory data shows no clear trend among the participants between the two different scenarios. A lowered breath rate is described to be an indicator of a higher perceived workload [35]. However, only participants 4, 7, and 13 showed clearly distinguishable

signals during the manual task, as shown in Fig. 7 part C. This result is rather unexpected as respiration is described to be strongly linked with ECG data [13, 28]. Yet, overall, the ECG data showed a more consistent trend among different participants. Consequently, low machine learning prediction accuracies would be expected based on the acquired respiration data.

Regarding heart rate and breath rate, it needs to be considered that the observation window, in this experiment, was between 36 and 154 s. This is fairly short, considering ECG measurements for medical applications are typically conducted over several hours [13, 28]. Therefore, it is expected that a longer observation period would lead to more distinguishable results.

Rather surprising results were achieved, at first, in the follow-up questionnaire (NASA-TLX) after the experiment, shown in Fig. 8 part A. The EEG and ECG data indicated higher workloads during the manual task and lower workloads during the collaborative task. In total, 11 out of 18 participants confirmed this trend in their subjective observation. However, participant 2 described the perceived workload as considerably higher during the collaborative scenario. Yet, there are no indications in the physiological data to back up
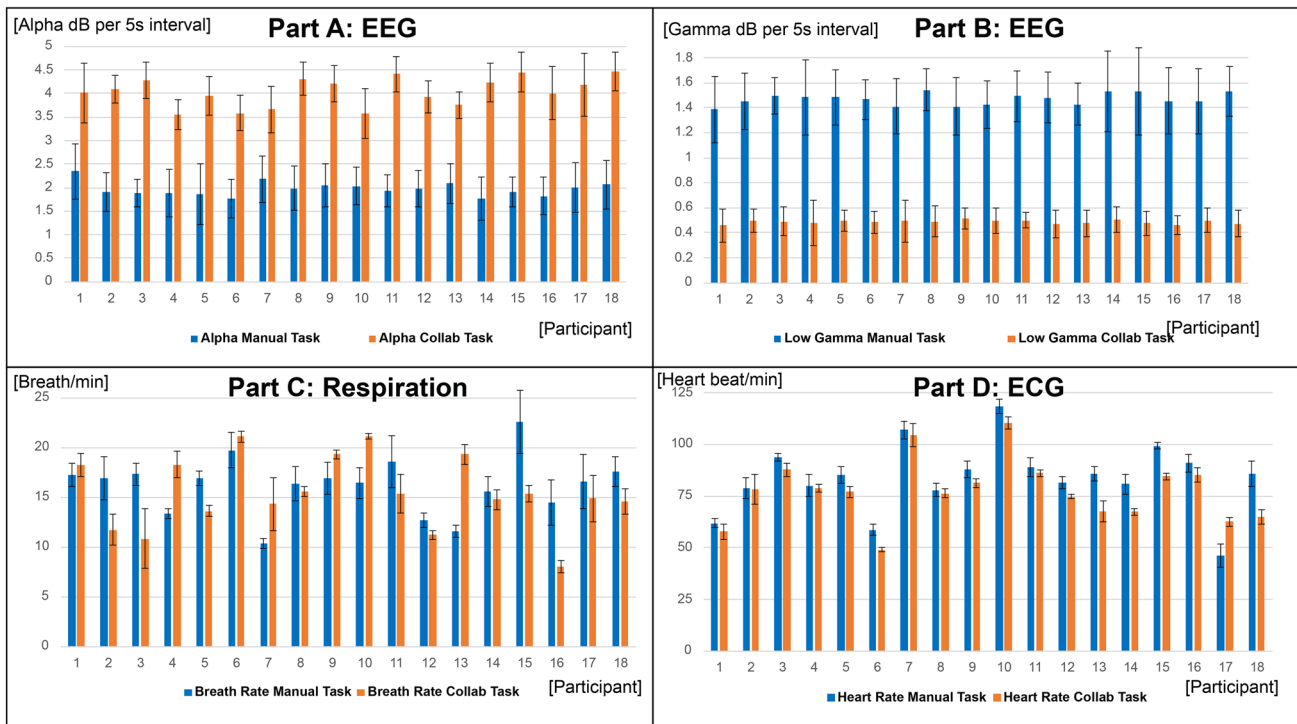
Fig. 7 Physiological results

this observation. The participant suggested that the subjective high perceived workload might be due to a low trust in the system/robot.

Participants 1, 7, 9, 15, and 18 described the perceived workload as equal during both scenarios. This would lead to two possible conclusions: either participants have a poor judgement of their perceived workload in retrospect, as one potential explanation [43]; or that the physiological sensors did not capture all relevant features to correctly predict those participant's perceived workload. Further analysis showed that the second conclusion points towards being correct. Overall, the NASA-TLX incorporates physical demand, mental demand, temporal demand, effort, performance, and frustration [42]. These metrics are aggregated to predict perceived workload. However, the individual metrics showed a larger degree of variation between the manual and collaborative scenario, as shown in Fig. 8 parts B and C (the performance rating is inverted).

Participants often rated the physical demand, mental demand, and effort higher during the manual task, as shown in Fig. 8 parts B and C. Though, the reason why the overall perceived workload is similar, for some participants, is due to their high rating of either temporal demand or frustration during the collaborative scenario. The participants stated this was either due to the robot moving too fast or too slowly. Further analysis and experiments could investigate how a

collaborative robot's speed affects the overall perceived workload of the human operator. Moreover, this highlights the potential for an adapted robot behavior, which adjusts its speed based on the preference and working pace of the human operator.

Overall, the average subjective performance rating was higher (inverted in figure) during the collaborative scenario than during the manual task, which was also observed in the objective data (Fig. 9). The mean duration of the manual task was 58 s and 52 s for the collaborative task. Also, 6 mistakes and 9 assistance requests were recorded during the manual scenario. Only four assistance requests and no errors were observed during the collaborative task. Thus, this data points towards slightly higher performances during the collaborative scenario.

Demonstrating the adaptive ability of the framework through prediction of perceived workloads might not be fully achievable from this data, due to do some participant's higher rating of mental demand, physical demand, and effort during the collaborative scenario. This led to similar or equal perceived workload scores during the different scenarios, despite having different causes. Thus, to test the framework, an additional metric was established that considers individual scores of mental demand, physical demand, and effort (MPE).

In order to validate whether the framework can learn to predict either perceived workload or MPE from
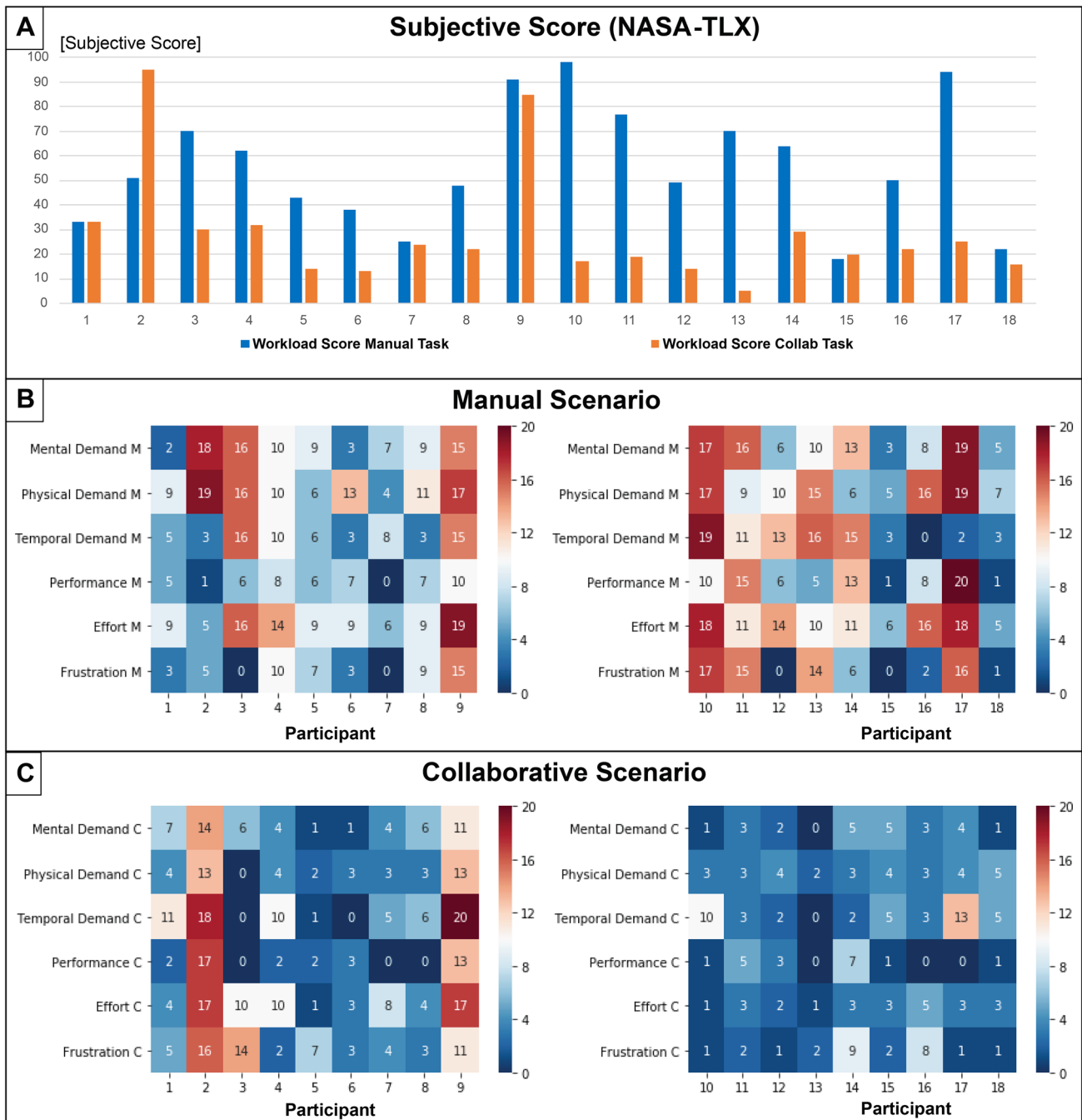
Fig. 8 Subjective results (TLX)

physiological data, a testing scenario was developed. For each participant, a sequence of five second intervals of the collected data was streamed into the framework, in the same sampling rate as it was acquired. This consisted of both the fully manual and the collaborative scenario. Each sensor was included in a service, which included preprocessing and predictions based on machine learning. For the EEG data,

the preprocessing consisted of a notch filter and fast Fourier transform. The resulting alpha and gamma waves were used to train a support vector machine (SVM). The ECG and respiration data were filtered based on the confidence level within the sensor data, which was provided by the device itself. For classifying sensor data, also an SVM was chosen with a Gaussian kernel in each service. Yet, other classifiers
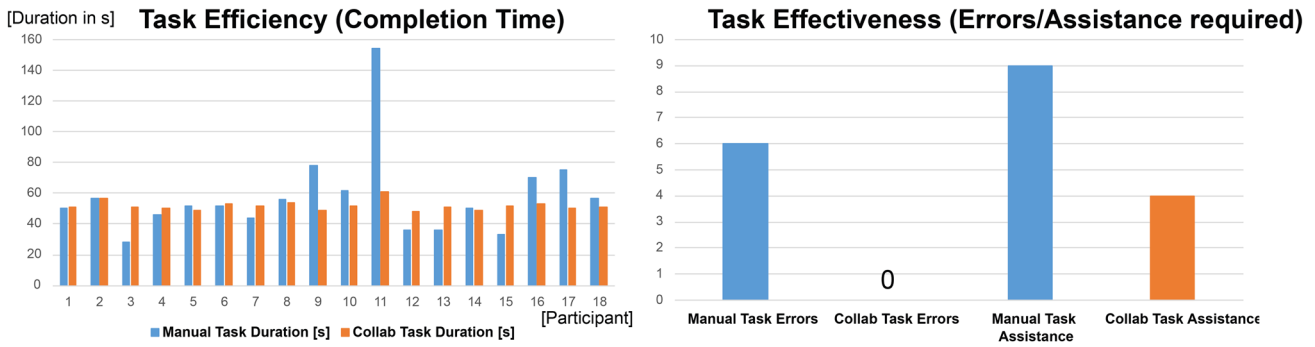
Fig. 9 Objective results

such as random forests and neural networks delivered similar results.

In a first step, the SVM was trained offline based on an 80/20 split, which means that 80% of the available data was used to train a classifier and 20% for testing. In the context of physiological data, however, it is important to not break the time domain of the data as it often follows the aforementioned sinusoidal pattern. Otherwise, unrealistic constellations (different wave patterns) might occur. Thus, instead of performing an 80/20 split on raw data, the split was performed based on the windowed, Fourier transformed intervals. Finally, the trained classifiers were included in a service, and the experimental data was streamed into this service in the same sampling rate as it was acquired, to test the online classification capabilities. After each 5-s interval, the prediction of a sensor/service is compared with the subjective data to adjust the services weights using reinforcement learning, in which a correct prediction increases the weight of a sensor. An incorrect prediction, on the other hand, will severely reduce the weight of a sensor, as a penalty function. Subsequently, based on the associated weight of a sensor, its suitability in the current context could be evaluated.

To test the frameworks' adaptive capabilities, two testing cases were implemented, based on the experimental data, shown in Figs. 7 and 8. On one hand, this included an obvious difference in perceived workload (participant 10). Whereas the participant indicated high perceived workloads during the manual scenario, and lower levels during the collaborative task, as expected. And, on the other hand, a minor, inverted difference in perceived workload (participant 15), which was not aligned with the expected outcome. Although participant 15's subjective rating could be influenced by cognitive bias, the participant was not confronted with the result. In fact, this constellation was considered an ideal testing scenario, to evaluate the framework's capability to raise awareness of different metrics not correlating. Figure 10 parts A and B show the comparison of the perceived workload prediction between participants 10 and 15.

Participant 10 had a perceived workload score of 98 during the manual task and 17 during the collaborative task. Subsequently, the weight of the EEG sensor is continuously increased, whereas the weight for breath rate is reaching a weight of zero. The heart rate is correlating, except for one 5-s interval. Based on these learning outcomes, the EEG and ECG would be considered the most suitable sensors in that context. Also, the framework would be able to distinguish the manual from the collaborative scenario, based on this data.

Participant 15 had a perceived workload score of 18 during the manual scenario and 20 during the collaborative task. Figure 10 part B shows that all physiological sensors were ineffective to predict the states. Thus, a consistent perceived workload prediction across all participants was not possible with the sensors used. However, this is mainly due to slightly higher temporal demands and frustration levels that were observed during the collaborative scenario. In contrast to perceived workloads, considering the individual subcategories, MPE showed a more consistent result among all participants. Despite the fact that both participants had different perceived workload scores, the framework could predict all 41 intervals correctly regarding MPE for participant 10; and all 31 instances correctly for participant 15. This demonstrates the framework's ability to compare its predictions with subjective measures and thus, to adjust the weights of sensors. In addition, the framework established an awareness of discrepancies between the subjective input and physiological metrics, as in the case of participant 15 for predicting perceived workloads.

Overall, these results highlight the complexity of measuring the human and establishing common interpretations among different individuals. Despite participants completing the same tasks, their subjective and physiological responses showed large variations, as shown in Figs. 7 and 8. Nevertheless, this can also be seen as an opportunity. Since participant's perceived workloads are described to affect task performance, adapting to the unique preferences of a human operator could enable a higher system's performance. Thus,

Fig. 10 Framework learning results
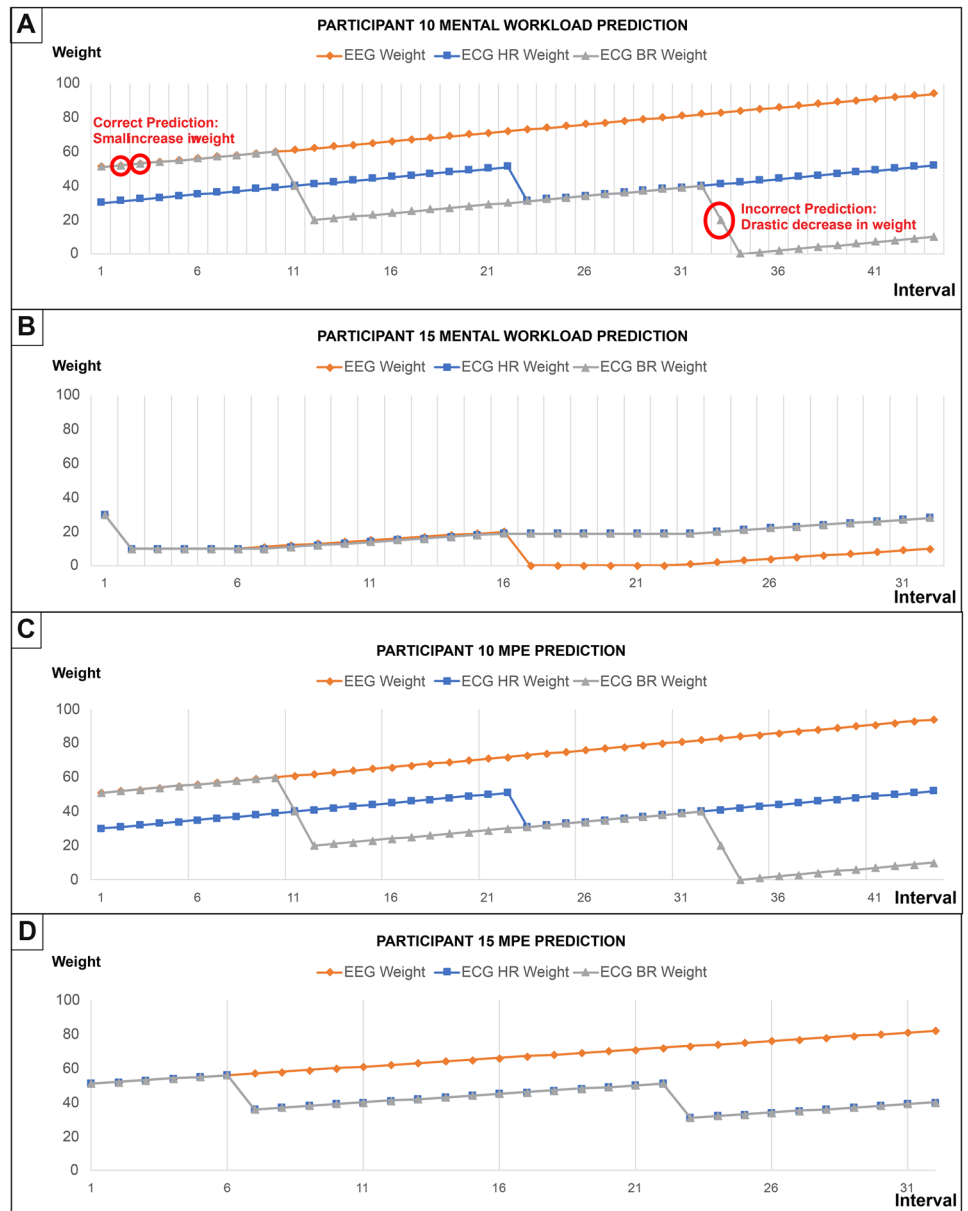
in a next step, the framework, including an agent-based task allocation/organization, will be applied online. This would allow for validating whether the proposed framework enables a higher productivity than existing systems.

6 Conclusion and future work

A novel adaptive sensor framework was introduced to establish data-driven models of the human for HRC. The framework incorporates subjective, objective, and physiological metrics in conjunction with machine learning to process sensory data, as well as to perform reinforcement learning, based on the operator's subjective input. By giving the

framework the ability to learn and adjust itself, it offers the potential to reduce manual fine-tuning efforts when establishing data-driven models. Since the framework offers a plug-and-play architecture, it could be applied to predict various physiological states of the human operator, with different wearable sensors, depending on the context. Thus, it enables various potential use cases of implementation/data-driven models according to Marr's framework in HRC.

To test the framework's adaptive behavior, it was being used during two assembly scenarios (manual and collaborative) to predict perceived workloads of the human operator. Perceived workloads are described to have a substantial impact on human task performance. Although the scenarios were expected to provide clear and distinguishable results

(high workloads during manual scenario, low workloads during collab scenario), this did not apply across all participants. Some participants rated the perceived workloads equally, due to the collaborative robot either moving too fast and causing high temporal demands or the robot moving too slowly which caused frustration. Thus, further research could investigate how adaptive robot speeds could affect the human operator's performance.

In addition, participants showed variation in both physiological signals and subjective perception, despite completing the same tasks, which highlights the uniqueness of human beings and their subsequent distinct behavior. By adapting to the individuality and the dynamic nature of human states and behavior, data-driven models of the human operator are expected to provide a more accurate representation of the human goal-directed behavior. This is due to the use of subject-specific, real-time measured values that are interpreted by machine learning models. Allowing the robot to adapt and adjust its behavior is envisioned to enable a higher system's performance.

Overall, the framework depicted in Fig. 4 could be further expanded in mainly two ways, namely, on the horizontal scale and the vertical scale.

On a horizontal scale, a wider range of physiological sensors can be deployed to more accurately predict the state of the human operator. As such, this could include pupillometry or skin conductance activity, among others. Moreover, two data services originating from one sensor device are also plausible. For instance, an ECG could provide both heart rate and heart rate variability. Therefore, this would allow for comparing different physiological metrics, even within the same sensor data.

On a vertical scale, a sensor data stream could be interpreted by different machine learning algorithms. Thus, this would allow for comparing the performance of more simple classifiers, such as SVMs with advanced deep learners such as LSTM-RNN's. In addition, the framework could be utilized to predict a wider set of human states, such as the physical state for example. For this purpose, EMGs could be deployed to predict the muscular fatigue of a human operator. In this case, a collaborative robot could assist the operator during the manipulation of heavy payloads, to establish a more ergonomic environment.

In a next step, the framework including the human-agent incorporating a data-driven model for perceived workloads will be applied online. Thus, it would allow for quantifying the potential benefit of an adaptive behavior over existing solutions regarding task allocation/organization.

In conclusion, HRC still offers a large unexplored potential, especially regarding the integration of the human. Advances in sensor technology in conjunction with machine learning offer promising potentials to further push boundaries towards a more synergetic collaboration between human operators and robots.

Declarations

Conflict of interest The authors declare no competing interests.

Availability of data and material Not applicable.

Code availability Code is available when required.

References

1. Mabkhot MM, Al-Ahmari AM, Salah B, Alkhalefah H (2018) "Requirements of the smart factory system: a survey and perspective," Machines, vol. 6, no. 2

2. Lafley AG, Charan R (2008) The Game-Changer: How You Can Drive Revenue and Profit Growth with Innovation. Crown Publishing Group

3. Wang L et al (2019) "Symbiotic human-robot collaborative assembly," CIRP Ann

4. Müller R, Vette M, Scholer M (2016) Robot workmate: a trustworthy coworker for the continuous automotive assembly line and its implementation. Procedia CIRP 44:263–268

5. Lasota PA, Shah JA (2015) Analyzing the effects of human-aware motion planning on close-proximity human-robot collaboration. Hum Factors 57(1):21–33

6. Pérez L, Rodríguez-Jiménez S, Rodríguez N, Usamentiaga R, García DF, Wang L (2020) Symbiotic human–robot collaborative approach for increased productivity and enhanced safety in the aerospace manufacturing industry. Int J Adv Manuf Technol 106(3–4):851–863

7. Robla-Gomez S, Becerra VM, Llata JR, Gonzalez-Sarabia E, Torre-Ferrero C, Perez-Oria J (2017) Working together: a review on safe human-robot collaboration in industrial environments. IEEE Access 5:26754–26773

8. Bruno G, Antonelli D (2018) Dynamic task classification and assignment for the management of human-robot collaborative teams in workcells. Int J Adv Manuf Technol 98(9–12):2415–2427

9. Hiatt LM, Narber, C, Bekele, E, Khemlani SS, Trafton JG (2017) "Human modeling for human–robot collaboration," Int J Rob Res 36(5–7):580–596.

10. Papanastasiou S et al (2019) Towards seamless human robot collaboration: integrating multimodal interaction. Int J Adv Manuf Technol 105(9):3881–3897

11. Wang L, Mohammed A, Wang XV, Schmidt B (2018) Energy-efficient robot applications towards sustainable manufacturing. Int J Comput Integr Manuf 31(8):692–700

12. Faccio M, Minto R, Rosati G, Bottin M (2020) The influence of the product characteristics on human-robot collaboration: a model for the performance of collaborative robotic assembly. Int J Adv Manuf Technol 106(5–6):2317–2331

13. Young MS, Brookhuis KA, Wickens CD, Hancock PA (2015) "State of science: mental workload in ergonomics," Ergonomics vol. 58, no. 1. Taylor & Francis, pp. 1–17

14. Perugini M, Conner M (2000) Predicting and understanding behavioral volitions: the interplay between goals and behaviors. Eur J Soc Psychol 30(5):705–731

15. Shiu EMK, Hassan LM, Thomson JA, Shaw D (2007) An empirical examination of the extended model of goal-directed behaviour: assessing the role of behavioural desire. Eur Adv Consum Res 8(January):66–71

16. Michie S, van Stralen MM, West R (2011) "The behaviour change wheel: a new method for characterising and designing behaviour change interventions," Implement Sci vol. 6, no. 1

17. Ritchie JB (2019) The content of Marr's information-processing framework. Philos Psychol 32(7):1078–1099

18. Pellegrinelli S, Moro FL, Pedrocchi N, Molinari Tosatti L, Tolio T (2016) "A probabilistic approach to workspace sharing for human–robot cooperation in assembly tasks," CIRP Ann. - Manuf. Technol vol. 65, no. 1, pp. 57–60

19. Breazeal C, Gray J, Berlin M (2009) An embodied cognition approach to mindreading skills for socially intelligent robots. Int J Rob Res 28(5):656–680

20. Carberry S (2001) Techniques for plan recognition. User Model User-adapt Interact 11(1–2):31–48

21. Peternel L, Tsagarakis N, Caldwell D, Ajoudani A (2018) Robot adaptation to human physical fatigue in human–robot co-manipulation. Auton Robots 42(5):1011–1021

22. Buerkle A, Al-Yacoub A, Ferreira P (2020) "An incremental learning approach for physical human-robot collaboration," in TAROS 2020 LNAI 12228 pp. 308–313

23. Al-Yacoub A et al (2021) "Data-driven modelling of human-human co-manipulation using force and muscle surface electromyogram activities," Electronics vol. 10, no. 13

24. Buerkle A, Bamber T, Lohse N, Ferreira P (2021) Feasibility of detecting potential emergencies in symbiotic human-robot collaboration with a mobile EEG. Robot Comput Integr Manuf 72(December):2021

25. Buerkle A, Eaton W, Lohse N, Bamber T, Ferreira P (2021) "EEG based arm movement intention recognition towards enhanced safety in symbiotic human-robot collaboration," Robot. Comput. Integr. Manuf vol. 70, no. February

26. "Health and Safety Executive" (2018) Managing Shiftwork, [Online]. Available: http://www.hse.gov.uk/pUbns/. [Accessed: 26-Jun-2019]

27. Steinfeld A et al (2006) "Common metrics for human-robot interaction," Assoc Comput Mach

28. Lazar J, Feng JH, Hochheiser H (2017) "Measuring the human," in Research Methods in Human-Computer Interaction, 2nd Eds. Boston: Morgan Kaufmann pp. 369–409

29. Mukhopadhyay SC (2015) Wearable sensors for human activity monitoring: a review. IEEE Sens J 15(3):1321–1330

30. Subhani, A. R., Xia, L., & Malik, A. S. (2011). EEG signals to measure mental stress. 2nd International Conference on Behavioral, Cognitive and Psychological Sciences, 84–88.

31. Minguillon J, Lopez-Gordo MA, Pelayo F (2016) "Stress assessment by prefrontal relative gamma," Frontiers in Computational Neuroscience 10:101

32. Duchowski AT (2017) Eye Tracking Methodology

33. Cho Y, Bianchi-Berthouze N, Oliveira M, Holloway C, Julier S (2019) "Nose heat: exploring stress-induced nasal thermal variability through mobile thermal imaging", 2019 8th Int. Conf Affect Comput Intell Interact ACII 2019:566–572

34. Abdelrahman Y, Velloso E, Dingler T, Schmidt A, Vetere F (2017) Cognitive heat: exploring the usage of thermal imaging to unobtrusively estimate cognitive load. Proc ACM Interactive, Mobile, Wearable Ubiquitous Technol 1(3):1–20

35. Castor M et al (2003) "GARTEUR handbook of mental workload measurement," Group no. July, p. 164

36. Tan JTC, Duan F, Zhang Y, Watanabe K, Kato R, Arai T (2009) "Human-robot collaboration in cellular manufacturing: design and development", 2009 IEEE/RSJ Int. Conf Intell Robot Syst IROS 2009:29–34

37. Al-Yacoub A, Buerkle A, Flanagan M, Ferreira P, Hubbard E, Lohse N (2020) "Effective human-robot collaboration through wearable sensors," in Emerging Technologies and Factory Automation

38. Peternel L, Fang C, Tsagarakis N, Ajoudani A (2019) A selective muscle fatigue management approach to ergonomic human-robot co-manipulation. Robot Comput Integr Manuf 58(January):69–79

39. Ferreira P, Lohse N, Ratchev S (2009) Multi-agent architecture for self-configuring modular assembly systems. IFAC 42:16

40. Liu Z, Liu Q, Wang L, Xu W, Zhou Z (2021) "Task-level decision-making for dynamic and stochastic human-robot collaboration based on dual agents deep reinforcement learning," Int J Adv Manuf Technol 3533–3552

41. Gawron VJ (2008) Human Performance, Workload, and Situational Awareness Measures Handbook. CRC Press

42. NASA (1968) "Task Load Index," 1–26

43. Jahedi S, Méndez F (2014) On the advantages and disadvantages of subjective measures. J Econ Behav Organ 98:97–114

44. "OPC UA" (2020) [Online]. Available: https://opcfoundation.org/about/opc-technologies/opc-ua/. [Accessed: 19-Oct-2020]

45. Profanter S, Tekat A, Dorofeev K, Rickert M, Knoll A (2019) "OPC UA versus ROS, DDS, and MQTT: Performance evaluation of industry 4.0 protocols," Proc IEEE Int Conf Ind Technol 955–962

46. Bi L, Feleke A, Guan C (2019) A review on EMG-based motor intention prediction of continuous human upper limb motion for human-robot collaboration. Biomed Signal Process Control 51:113–127

47. Bouchachia A, Gabrys B, Sahel Z (2007) "Overview of some incremental learning algorithms," IEEE Int Conf Fuzzy Syst