Measuring Human Comfort in Human-Robot Collaboration via Wearable Sensing

Yuchen Yan, Haotian Su and Yunyi Jia

Abstract— The development of collaborative robots has enabled a safer and more efficient human-robot collaboration (HRC) manufacturing environment. Tremendous research efforts have been conducted to improve user safety and robot working efficiency after the debut of collaborative robots. However, human comfort in HRC scenarios has not been thoroughly discussed but is critically important to the user acceptance of collaborative robots. Previous studies mostly utilize the subjective rating method to evaluate how human comfort varies as one robot factor changes, yet such method is limited in evaluating comfort online. Some other studies leverage wearable sensors to collect physiological signals to detect human emotions, but few of them implement this for a human comfort model in HRC scenarios. In this study, we designed an online comfort model for human-robot collaboration using wearable sensing data. The model uses physiological signals acquired from wearable sensing and calculates the in-situ human comfort levels based on our developed algorithms. We have conducted experiments in realistic human-robot collaboration tasks, and the prediction results demonstrated the effectiveness of the proposed approach in identifying human comfort levels in HRC.

Index Terms—Wearable Sensing, Human Comfort, Human-robot Collaboration

I. Introduction

The Human-Robot Collaboration (HRC), known as "the state of a purposely designed robotic system and operator working in a collaborative workspace" [1], has gained growing attention in its research field during the past few years. However, the market share and industry-level applications of these collaborative robots (COBOTs) are still limited and have huge space for improvement. One of the customers' concerns for these COBOTs originates from user acceptance, which is highly influenced by the perceived human comfort of the worker. The comfort of human plays such a critical role that not only does it affect the user acceptance but also has a significant impact on the efficiency of manufacturing, which has become a critical issue [2-4]. For example, Ye et al. [4] found that workers' performance varied significantly under different thermal comfort conditions. The productivity would decrease by 9% when the temperature changes from 25.0 to 35.0 °C.

Prior to discussing any human comfort evaluation method or theory, the concept of comfort and some basic background knowledge need to be clarified and introduced first. The disappointing fact is that the academia has not come to a consensus on a universal definition of comfort yet, thus it still remains a huge challenge to precisely evaluate human comfort level [5]. Some researchers perceived comfort as two discrete states: comfort presence and comfort absence, where comfort has been simply defined as the absence of discomfort and vice versa, while some others held the contrasting opinion which claims that comfort and discomfort are two opposites on a continuous scale, ranging from extreme discomfort through a neutral state to extreme comfort [6][7]. Some researchers also viewed comfort as an optimal state in which the person stops taking actions to avoid discomfort [8]. Despite all the arguments and disagreements in academia, people come to a common agreement on several points: (1) comfort is subjectively determined by each individual's personal nature; (2) comfort can be affected by a wide variety of factors from multiple natures such as physical, physiological or psychological; and (3) comfort is affected by one's reaction to the environment stimulus [6]. These statements were also used as the guidelines in our study.

In recent years, some research efforts have been spent on human comfort evaluation and adaptation in HRC manufacturing tasks. For instance, Weitian. et al. [9] proposed a computational approach to model and quantify the human comfort during human-robot collaborative manufacturing. Ross. et al. [10] found that human comfort has a direct and immediate influence on the collaboration quality between the robot and its human partner, is also a significant factor for the robot to be aware of. Jessi. et al. [11] developed a method of evaluating how the invasion of personal space by a robot affects human comfort. Przemyslaw. et al. [12] examined human response to motion-level robot adaptation to determine its effect on team fluency, human satisfaction, and perceived safety and comfort. Alami et al. [13] proposed a framework that allows the robot to select and perform its tasks based on the human partner's presence, needs, and preferences. Ciccarelli et al. [14] proposed a system to improve human postural comfort by optimizing robot behavior.

However, most of these current research methods on human comfort in HRC tasks merely utilize subjective ratings or simple statistical comparison approaches. Thus, the results of the papers above can only prove the qualitative or simple quantitative relationship between human comfort levels and the HRC factors. Limited research has fully leveraged the advantage of comfort measurements by utilizing physiological

Yuchen Yan, Haotian Su, Yunyi Jia are with the Department of Automotive Engineering and International Center for Automotive research at Clemson University, Greenville, SC 29607 USA. (e-mail: yucheny@clemson.edu, haotias@clemson.edu, and yunyij@clemson.edu). Corresponding author: Yunyi Jia.

signals, e.g., electroencephalography (EEG), electrodermal activity (EDA), blood volume pulse (BVP), in a machine-learning-based model to analyze the general human comfort in HRC.

Some studies in the psychological field have already proved the effectiveness and feasibility of implementing machine learning-based or neural network-based methods to evaluate human mental activities such as cognitive load and emotion states, with either single type of physiological signal or combined features from multiple types of signals [15][16]. Shan. et al. [17] applied machine learning techniques in conjunction with passive EEG measurement to classify occupants' real-time thermal comfort states. Performances of different machine learning techniques were compared, and methods to select linear continuous features for class interpolation were also explored. The classification results with the linear discriminant analysis classifier using the full-set features achieved an accuracy above 90%. Maaoui's work [18] used two methods, support vector machine (SVM) and Fisher discriminant, to recognize human emotions of amusement, contentment, disgust, fear, neutral, and sadness with multiple physiological signals, e.g., BVP, EDA, Skin Temperature (SKT). Recognition results for different types of emotions turned out to be excellent with the accuracy around 92%. Kang. et al. [19] studied visual discomfort by applying the SVM approach and built a brain-computer interface framework to optimize the stereoscopic 3D content based on the viewer's EEG response.

In summary, despite a great amount of research efforts have been put into human comfort and physiological signal studies, two major research gaps remain today. The first gap is the lack of an approach to build an individual-based human comfort model that can accurately predict human comfort. The second gap is the lack of explorations of utilizing physiological signals in measuring overall human comfort in HRC scenarios, since most studies only focus on one specific feature. One thing worth noting is that the exact definition of comfort is still under debate within academia, since it is still considered as a highly subjective feeling, which can not be simply considered the same as stress or emotion. In this study, the motivation is to develop an AI-model-based framework which predicts general human comfort levels during HRC tasks based on physiological signals and potentially use this model and comfort data to optimize human comfort during future HRC tasks. The performance of our model is determined by comparing the differences between the physiological prediction model results and the human-reported Likert Scale ratings, which are used as the ground truth. A series of HRC tasks with five varying robot-motion factors were designed and used in the experiment. We implemented two comfort measurement approaches in our study – the subjective method and the objective method. Two types of data, subjective comfort ratings based on a Likert Scale and objective physiological signals, were collected online in this study as the experiment progressed. Then, we tested the effectiveness of our developed model which uses multiple machine learning/deep learning techniques by training and testing the model with the data we collected. Previous studies mostly adopt SVM-based feature extraction method only, we introduced and validated three types of feature extraction methods, including the autoencoder-based method, which has been rarely applied in physiological signal-based human factor studies before. Since the data used for the comfort model is physiological signal during human-robot collaboration, the solution is independent of human-robot collaboration tasks and scalable to the application of detecting human comfort levels in any physical human-robot collaboration contexts. In addition, the comfort models are built based on individual-dependent physiological data instead of mixing all subjects' comfort data for model training.

II. Experiment and Data Acquisition

A. Experimental Platform

The experimental platform is shown in Fig. 1. The collaborative robot used in this study is an ABB-14000 YUMI model. The Yumi robot is installed and centered on the backside of the black experiment platform, while two small cubes which are used for interaction tasks are placed on two farther corners of the platform respectively. The test subject will be standing in front of the experiment platform with a horizontal distance of 20cm. The Yumi robot is controlled by our built control system in ROS [20]. The higher-level YUMI motions for both arms are generated and executed in ROS.

B. Human-Robot Collaboration Tasks Design

In this study, we adopted a simple robot-delivery action as the interactive task. In total, we created 58 robot-delivery tasks, each consisting of a unique combination of factor levels. As shown in Table 1, there are totally five factors used in our study, while four of them are robot motion-based factors such as robot moving speed, final delivery distance, final delivery height and delay/waiting time. Delay period refers to the time length of the stagnation between robot's pickup and delivery actions. Robot speed refers to the linear moving speed of the robot tool center point (TCP). Final delivery distance refers to the shortest horizontal distance between the tip of the robot TCP and the human subject. Final delivery height refers to the vertical distance between the robot's TCP and the working platform. The fifth factor is unique in our study, which is the Left/Right Working Arm of Yumi. This is enabled by the unique double-arm design of Yumi [21]. Different robot arm selection will affect the selection of human arm for interaction by the human subject, we believe such differences could also induce human comfort variation.

As shown in Table 1 below, each factor has seven levels to choose from, except for robot arm selection has only left/right options. Different values for each factor were chosen and combined into a factor set which forms one experiment task. Each task only tunes one robot motion factor at a time, while keeping other motion factors at their medium levels. For each formed combination set, there are two mirrored scenarios generated by the left and right arms used in the task, which doubles the total number of tasks. There are also two extra reference cases which take the medium levels from all factors, one case for the left arm, the other one for the right arm. Eventually, 58 combination sets were created, and each one was used as the robot motion planner inputs for the task. The focus

of the experimental designs is to generate the physiological signals of humans under different comfort levels during human-robot collaboration. Thus, the tasks we designed are sufficient to generate enough physiological data under different comfort levels to conduct the training and testing of our proposed approaches. This will benefit a wide range of human-robot collaboration tasks in various manufacturing contexts.

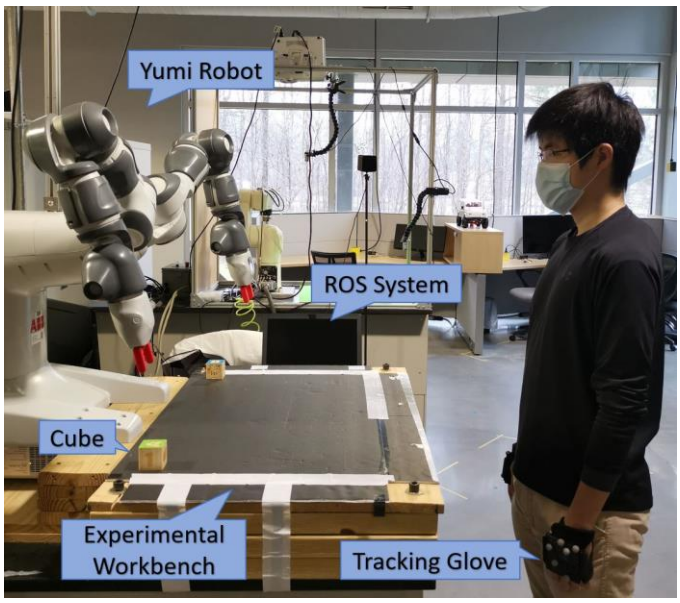

Fig. 1. Pressing Force Collection Device and its Diagram

TABLE1. THE FACTOR COMBINATION SET TABLE

Factors / Levels	Delay Period (s)	Robot Speed	Distance (cm)	Height (cm)	Robot Arm
1	0	0.1	25	15	Left / Right
2	1	0.2	35	20	
3	2	0.3	45	25	
4	3	0.4	55	30	
5	4	0.5	65	35	
6	5	0.6	75	40	
7	6	0.7	85	45	

C. *Subjective Comfort Level Acquisition*

To evaluate the subjective comfort levels of the participants for each HRC task, a 5-point Likert Scale was used in this study. After completing each HRC task, the participant would report a score scaling from 1 to 5 as the comfort level evaluation feedback. A score of 5 indicates that the test subject feels completely comfortable, while a score of 1 indicates the subject feels completely uncomfortable. These subjective ratings were the ground truth labels for the training process of the machine learning model. Training sessions were carried out before the official experiment started in order to let test subjects get a rough concept of what the extreme condition scenarios feel like. Also, test subjects will be instructed to try their best ignoring any other factors that they found distracting or irrelevant to the experiment design.

D. *Physiological Data Collection*

Physiological data collected in this study include EDA, heart rate (HR), BVP, SKT, and EEG. Two wearable sensing devices were used in this study. The first device, Empatica E4 wristband, was used to measure EDA, HR, BVP and skin temperature signals, and the second device, Emotiv EpocX headset, was used for EEG signal collection.

EEG signals provide us with useful information in analyzing the high-level emotions of the test subjects [22]. The portable EEG device used in this study is the Emotiv EpocX headset. As shown in Fig. 2, it is equipped with 16 non-invasive electrodes which touch against a person's scalp to measure the electric potential values at corresponding locations. Note that there are two reference electrodes which do not directly provide EEG data, but are only used as the "ground." The EmotivPro Software, developed by the Emotiv Epoc Manufacturer, is integrated with online EEG data monitoring, data post-processing and high-level feature extraction functions. In this study, we used the EmotivPro for data recording and results exporting. The output measurement results include raw EEG data from 14 channels, frequency domain analysis data, and high-level emotion states, e.g., excitement, stress, and focus. We took advantage of the high-level emotion extraction function of the software and utilized the excitement and stress performance data [34] in our later analysis since excitement reflects the positive psychological and physiological arousal of the human body, while stress reflects the negative human reactions to the environment. Both signals have 0.1-Hz sampling rate and require interpolation during the data preprocessing stage. Details about data preprocessing will be introduced in a later section.

Fig. 2. Wearable Sensing Devices

EDA, also known as electrodermal activity, is the property of the human body that causes continuous variation in the electrical characteristics of the skin. Skin resistance varies with the state of the sweat glands in the skin. The arousal of the sympathetic autonomic nervous system activity can result in the increase of sweat gland, which leads to greater skin conductance. Thus, the EDA signal is widely used as another important index in evaluating a person's psychological or physiological arousal in response to an external stimulus [23]. The EDA signal collection device we used in this study is the Empatica E4 wristband, which is equipped with two AgCl plated electrodes on the strap. During the experiment process, the AgCl electrodes firmly touch against the skin of the inside

of the participant's wrist, in the meantime, the wristband passes a minuscule amount of current between two electrodes in contact with the skin, thus obtains the skin resistance values. The data was measured from test subject's non-dominant hand with a 4-Hz sampling rate.

BVP, which stands for blood volume pulse, measures heart rate based on the volume of blood that passes through the tissues in a localized area with each beat of the heart. BVP measurement is achieved with the photoplethysmography (PPG) sensor embedded in the Empatica E4 wristband. This component measures changes in blood volume in the arteries and capillaries that correspond to changes in the heart rate and blood flow. The sampling frequency of the BVP data is 64Hz.

SKT measures the thermal changes on the skin. Variations in SKT mainly result from localized changes in blood flow caused by vascular resistance or arterial blood pressure. Local vascular resistance is modulated by smooth muscle tone, which is mediated by the sympathetic nervous system. The SKT variation reflects autonomic nervous system activities and is an indicator of a person's psychological state [24]. The SKT had a 4-Hz sampling rate.

E. Experiment Procedure

As mentioned earlier, we created 58 robot delivery tasks based on the five factors. Each delivery task lasts between 18-25 seconds, depending on the selected robot speed and stagnation time. The general procedure of each task in concise is that the robot picks up a cube first, either from the left or right-hand side, and delivers it to the participant, then the participant takes the cube from the robot arm and reports the subjective comfort level rating for the finished task.

Fifteen healthy (thirteen males and two females) with a mean age of 27.7 (SD = 3.68) years old graduate students participated in the experiment. All participants had engineering backgrounds. Before the experiment, the participants were introduced to the experimental protocols and signed on the consent form for taking part in the study. After that, the experimenter would help them put on the physiological measurement devices. Then, the procedures of the experiment and the tasks for the participants were introduced in detail to them. Before the actual experiment process started, participants first undertook a thorough training session familiarizing themselves with Yumi and its delivering actions with all the sensing devices on their bodies. These training tasks are highly similar to the HRC tasks in the actual experiment. Participants should get fully accustomed to the feeling with all the wearable sensors on their bodies before starting the actual experiment. Several extreme condition scenarios with the highest levels of factors would be experienced by participants during the training session, and participants would be instructed to ignore any other factors that they found distracting or irrelevant to the experiment design. This training session was repeated until participants announced well-prepared for the experiment.

During the official experiment, the order of executing 58 HRC scenarios was shuffled to guarantee that the participant would not be affected by learning effects. All 58 scenarios were executed one by one in a non-stopping manner. The entire experiment process took approximately 20 minutes. We did not set up a break time for the test subjects due to the concern that a break time might break the consistency in the subject's judgment for comfort. Besides, none of our subjects claimed that they felt exhausted without a break after the experiment finished.

F. Data Preprocessing

It is critical to implement preprocessing on the raw data obtained from the experiment before the feature extraction. The very first step was to clean the data, which means eliminating all the data recorded before the starting timestamp marker and after the ending timestamp marker. All the physiological data were then synchronized based on the starting marker and then divided into 58 pieces, corresponding to 58 scenario tasks respectively. Next, the data of each task was sliced into 2-second-long window samples. The window size, two seconds, was determined based on the characteristics of the physiological signals. A two-second duration is long enough to capture a clear phasic skin conductance response, EEG response and also BVP changes. Since each HRC task lasts around 18-25 seconds, each task is composed of 9-13 serialized small samples, as shown in Fig. 3.

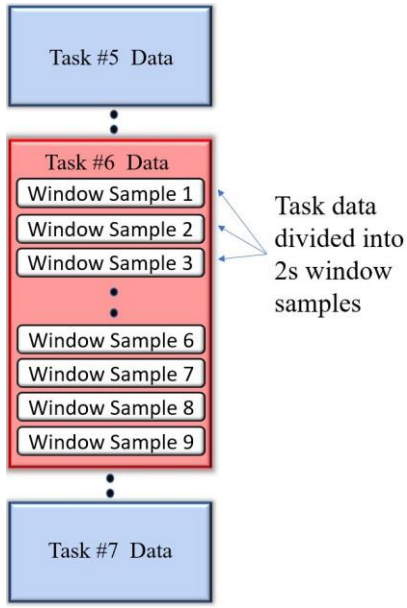

Fig. 3. Data Preprocessing and Window Sample Slicing

Skin conductance is typically characterized into two types – tonic skin conductance level (SCL, also known as the tonic component) and phasic skin conductance response (SCR, also known as the phasic component) [25]. Both components are widely used in most recent emotional change detection studies [26][27].

Ledalab, a Matlab-based software, was used in this study for tonic and phasic components extraction from the raw EDA signal.

G. Feature Extraction

With all the preprocessed data available in our hands, one last step before the machine learning model training process was to extract valuable features from the data.

Feature extraction process was executed within each 2-second data sample. Normalization was performed on the data to mitigate the influence of individual differences in physiological signals. The collected physiological data were normalized based on the average value within the corresponding task. The average value of a certain task and the normalized data was calculated using the following formula:

$$\bar{x}_{avg} = \frac{\sum x_{com}}{n_{com}} \quad (1)$$

$$x_{norm} = \frac{x_{raw} - \bar{x}_{avg}}{\bar{x}_{avg}} \quad (2)$$

where x_{norm} is the normalized signal, x_{raw} is the raw signal, $\bar{x}_{avg}$ is the average value of the signal within all samples after noise filtering, x_{com} is the signal within samples after noise filtering, n_{com} is the number of samples within the 2-second window.

Raw physiological data itself does not possess much useful information without proper feature extractions. According to Picard's work [29], there are several statistical features which are effective in emotion detection. Besides the raw signal x, other statistical features include the mean value μ_x, the standard deviation σ_x, the mean of the absolute values of the first differences δ_x, and the mean of the absolute values of the second differences γ_x. We also add several other statistical features which include the maximal value max_x, the minimal value min_x, the odds of the minimal value over the maximal value $ratio_x$ [30], and the root mean square value M_x. All the features mentioned above were calculated within each 2-second window sample for all types of physiological signals based on their original raw data and normalized raw data. The calculation formulas are listed below:

$$\mu_x = \frac{1}{T}\sum_{t=1}^{T} X(t) = \bar{X}_t \quad (3)$$

$$\sigma_x = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(X(t) - \mu_x)^2} \quad (4)$$

$$\delta_x = \frac{1}{T-1}\sum_{t=1}^{T-1}|X(t+1) - X(t)| \quad (5)$$

$$\gamma_x = \frac{1}{T-2}\sum_{t=1}^{T-2}|X(t+2) - X(t)| \quad (6)$$

$$max_x = max(x) \quad (7)$$

$$min_x = min(x) \quad (8)$$

$$ratio_x = \frac{min(x)}{max(x)} \quad (9)$$

$$M_x = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(X(t))^2} \quad (10)$$

where $X(t)$ is the normalized raw signal x at sample t, while T is the total number of samples within the corresponding 2-second window. With seven sources of physiological signals, eight statistical features for each signal, including original and normalized data, a total of 112 features were extracted from each 2-second window.

The last step of feature extraction was to unify the dimensions of the datasets from each task. As mentioned in the previous section, each HRC task lasts between 18 to 25 seconds, making each task composed of 9 to 13 serialized window samples. Here we extended the dimensions of the datasets with fewer than 13 window samples. The solution is to duplicate the tail section of the data and append it to the end of the original dataset. For instance, for a task dataset composed of 10 window samples, we copied the last three window samples and attached them to the end of the original dataset.

The entire feature extraction stage is finished up to this point, with each feature vector of the task having a length of 1456, which is a product of 112 and 13.

III. Comfort Level Identification Method

The overall structure of our approach consists of four major components – raw data collection stage, physiological feature extraction stage, feature reduction / compression stage, and comfort level classification stage. The raw data is preprocessed first and implemented with the feature extraction step, generating a relatively large feature matrix. Then the feature reduction step is carried out to shrink the original feature matrix for higher classification efficiency and effectiveness. Lastly, the classification process is implemented for final training. The first two stages have been explained in the previous sections, the last two stages are explained in this section.

A. Support Vector Machine

Support vector machine (SVM) is a powerful supervised learning model which has been widely used for classification tasks in various fields. In this study, we used a multi-class error-correcting output codes (ECOC) SVM classifier which is composed of multiple binary SVM learners to solve the problem. The input of the classifier was a chronologically ordered array of compressed or reduced physiological features of all the window samples from an entire task, and the output of the classifier was the predicted label of comfort level of the task.

An ECOC model leverages multiple binary SVM classifiers by integrating a coding matrix design and a decoding scheme [28]. The coding design determines the classes that the binary learners train on, while the decoding scheme determines how the results of the binary classifiers are aggregated. An example of a coding matrix is shown below:

TABLE2. CODING MATRIX OF ECOC MODEL

	Learner 1	Learner 2	Learner 3
Class 1	1	1	0
Class 2	-1	0	1
Class 3	0	-1	-1

The above example classification problem has three classes and three binary learners. During the training session, for learner 1, class 1 is treated as the positive class and class 2 is

treated as the negative class, while all data from other classes is ignored by learner 1. Other learners are trained in similar ways. Based on the work from Escalera et al., in loss-weighted decoding [29], the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation. The decoding scheme formula is given below:

$$\hat{k} = \underset{k}{argmin} \frac{\sum_{l=1}^{L} |m_{kl}| g(m_{kl}, s_l)}{\sum_{l=1}^{L} |m_{kl}|} \quad (11)$$

$$g = \frac{\max(0, 1 - y_j s_j)}{2} \quad (12)$$

where m_{kj} is element$(k, j)$of the coding design matrix M (that is, the code corresponding to class k of binary learner j). s_j is the score of binary learner j for an observation. g is the binary loss function. $\hat{k}$ is the predicted class for the observation.
y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j.

B. Feature Selection & Compression Method

As mentioned in the previous section, the final feature vector after preprocessing has a length up to 1456, which is extremely long and unsuitable for the machine learning model training process. Besides, the amount of training data is very small for our experiment, which could easily lead to model underfitting. Therefore, it is critical to apply feature reduction or feature extraction before the official training process. In this study, three different techniques, including the SVM Recursive Feature Elimination (SVMRFE) algorithm, Autoencoder, and Independent Component Analysis (ICA), were implemented independently, and the results were compared.

SVMRFE can rank feature importance levels and remove relatively insignificant feature variables in order to achieve higher classification performance. Firstly, an SVM classifier is trained, and then the SVMRFE algorithm uses the weight magnitude as a ranking criterion, and computing and comparing the ranking criteria of all features to eliminate the lowest ranking features [30]. Then, the entire process is repeated iteratively to obtain the required number of features. The overall structure of SVMRFE is shown below:

Algorithm1 for SVM-RFE

Inputs: Training examples $\boldsymbol{X}_0 = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_k, \ldots \boldsymbol{x}_l]^T$, Class labels $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots \boldsymbol{y}_k, \ldots \boldsymbol{y}_l]^T$
Output: Feature ranked list r = []
1: Initialize subset of surviving features $\boldsymbol{s} = [1, 2, \ldots n]$
2: Initialize feature ranked list $\boldsymbol{r} = [f_1, f_2, \ldots f_n]^T$
3: **while** **s** not empty **do**
4: Restrict training examples $\boldsymbol{X} \leftarrow \boldsymbol{X}_0(:, \boldsymbol{s})$
5: Train the classifier $\alpha \leftarrow SVM\ train(\boldsymbol{X}, \boldsymbol{y})$
6: Compute the weight vector of dimension length (**s**)

$$\boldsymbol{w} = \sum_k \alpha_k y_k \boldsymbol{x}_k$$

7: Compute the ranking criteria $c_i = (w_i)^2$
8: Find the feature with the smallest ranking criterion $f \leftarrow argmin(c)$
9: Update feature ranked list $\boldsymbol{r} \leftarrow [\boldsymbol{s}(f), \boldsymbol{r}]$
10: Eliminate low ranking features $\boldsymbol{s} \leftarrow \boldsymbol{s}(1: f-1, f+1: length(\boldsymbol{s}))$
11: **end while**
12: **return** $\boldsymbol{r}$

Algorithm2 for Autoencoder

Inputs: Training examples $\boldsymbol{X}_0 = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_k, \ldots \boldsymbol{x}_l]^T$, Class labels $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots \boldsymbol{y}_k, \ldots \boldsymbol{y}_l]^T$
Output: Extracted feature list r = []
1: Initialize the dimension of the code layer
2: Train the autoencoder $\alpha \leftarrow AE\ train(\boldsymbol{X}_0, \boldsymbol{y})$
3: Encode the test samples with the trained encoder
4: Update the extracted feature list $\boldsymbol{r} \leftarrow encoder\_\alpha\ (\boldsymbol{X}_{test})$
5: **return** $\boldsymbol{r}$

Algorithm3 for ICA

Inputs: Training examples $\boldsymbol{X}_0 = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_k, \ldots \boldsymbol{x}_l]^T$, Class labels $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots \boldsymbol{y}_k, \ldots \boldsymbol{y}_l]^T$
Output: Extracted feature list r = []
1: Generate the ICA model with a data matrix $\boldsymbol{X}_{train} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_k, \ldots \boldsymbol{x}_l]^T$ with n rows of samples and p rows of features
2: Initialize a random p-by-q weight matrix $\boldsymbol{W}$
3: Objective function $obj_{func} \leftarrow$ minimizes $\sum g(\boldsymbol{X}_{train}\boldsymbol{W})$ where $g = \frac{1}{2}\log(cosh(2x))$ *is constrast function*
4: Train the ICA model with all training samples
5. Obtain the resulting feature list $r \leftarrow \boldsymbol{X}\boldsymbol{W}_{optimal}$
6: **return** $\boldsymbol{r}$

The second feature reduction/extraction method used in this study is the Autoencoder. Comprised of an encoder and a decoder, an autoencoder is an unsupervised artificial neural network that compresses and encodes data and reconstructs the data back from the reduced encoded representation to a representation that is as close to the original input as possible [32]. Due to this special working mechanism, the trained encoder works as a great feature extraction/reduction tool. In our study, autoencoders were trained for feature compression so that the feature vector length could be significantly reduced.

The third feature extraction/reduction method is the ICA. ICA is a technique that allows the separation of a mixture of signals into their different sources by assuming non-Gaussian signal distribution [33]. The ICA extracts the sources by exploring the independence underlying the measured data. Firstly, a data matrix $\boldsymbol{X}_{train}$ with n rows of samples and p rows of features is generated. Then initialize the p-by-q weight matrix $\boldsymbol{W}$. The objective function attempts to obtain a nearly orthonormal weight matrix and to minimize $\sum g(\boldsymbol{X}_{train}\boldsymbol{W})$ by using a standard limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) quasi-Newton optimizer, where $g = \frac{1}{2}\log(cosh(2x))$ is the contrast function. After training the model with all training samples, the resulting optimal feature list $\boldsymbol{X}\boldsymbol{W}_{optimal}$ is then obtained.

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2024.3383296

C. Prediction Accuracy Evaluation Method

The prediction accuracy calculation approach used in this paper is not 'zero-sum'-like. Instead, we applied a multi-level accuracy calculation method. The general equation for accuracy calculation is shown as below:

$$Accuracy = 1 - \frac{\left|x_{ground\ truth} - x_{predicted}\right|}{5 - 1} \quad (13)$$

This calculation method makes sense because a closer prediction to the ground truth result should have higher accuracy than those results which completely fall into the opposite side.

IV. Experiment Results and Analysis

In this study, we performed the feature selection/reduction algorithms on the data from 15 different test subjects. Two hyperparameters were tuned in this study, feature reduction/extraction methods and the number of reduced/extracted features. For all participants, all three types of feature reduction/extraction methods were used. As for the number of reduced/extracted features, we used four different options for each participant. The options were 25-features, 50-features, 100-features and 200-features. For each hyperparameter combination set, the prediction algorithm was repeatedly executed 15 times for better statistically reliable data.

The distribution of training samples and testing samples was 75% for training and 25% for testing. The datasets underwent a 10-fold cross-validation during the training process.

A. Results of Comfort Level Detection

Fig. 4-6 give us the prediction accuracy results of comfort level detection when three different feature reduction/extraction methods – SVMRFE, Autoencoder, ICA were used with four different choices on the number of extracted features – 25, 50, 100 and 200. The curves represent the average accuracy of each hyperparameter combination set for each participant. The results of SVMRFE/AE/ICA feature reduction methods are represented in Fig. 4, 5 & 6, respectively. Table 3-5 show the average, minimum and maximum testing accuracies of the classifier when trained with SVMRFE, Autoencoder-based and ICA feature reduction/extraction methods, respectively.

a. Results of Different Feature Extraction Methods

As shown in Fig.4-6, for 10 out of 15 participants, the Autoencoder-based feature extraction method yielded the highest personal average prediction accuracy, while 3 out of 15 participants achieved their highest personal average accuracy with the SVMRFE method and 2 participants obtained achieved highest accuracy with the ICA method. And by using the Autoencoder-based feature extraction method, the prediction of comfort level achieved over 75% personal average accuracy with 13 participants. By using the SVMRFE method, the prediction of comfort level achieved over 75% personal average accuracy with 9 participants. By using the ICA method, the prediction of comfort level achieved over 75% personal average accuracy with 7 participants.

As shown in Table3-5, among all the options of feature reduction/extraction methods tested in the study, the Autoencoder method achieved the highest total average accuracy – 76.68% across all participants, while both SVMRFE and Autoencoder methods yielded the highest maximum accuracy – 92.85% at the same time, and ICA method yielded the lowest minimum accuracy. The prediction results of subjects #2, 6 and 10 from a test run are shown in Fig. 7.

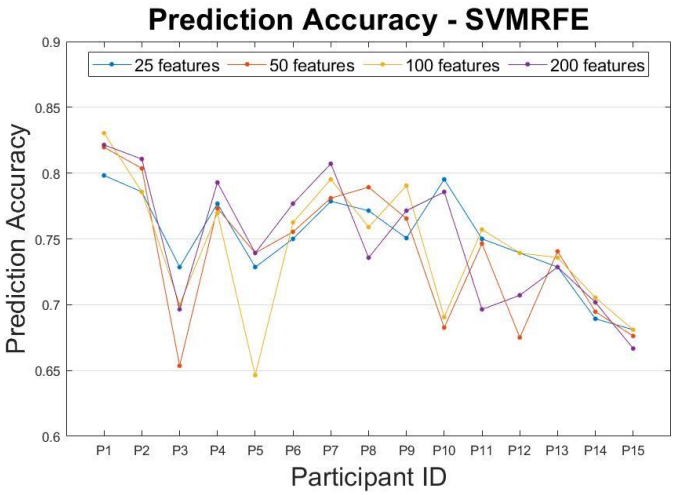

Fig. 4. Prediction Results of All Participants with SVMRFE Method

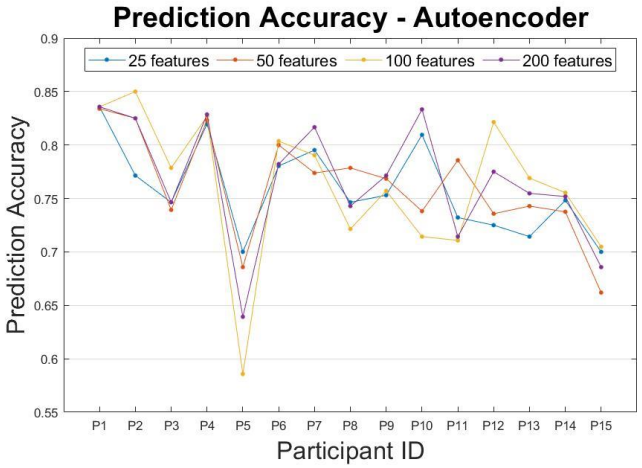

Fig. 5. Prediction Results of All Participants with Autoencoder Method

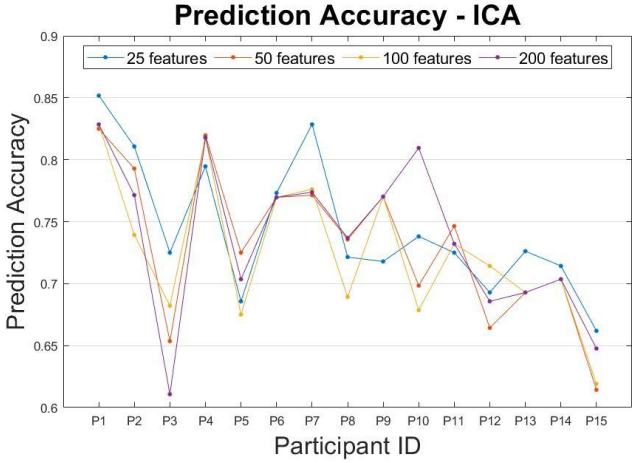

Fig. 6. Prediction Results of All Participants with ICA Method

TABLE3. TESTING ACCURACY OF THE SVMRFE FEATURE REDUCTION METHOD WITH 25-FEATURES, 50-FEATURES, 100-FEATURES, 200-FEATURES

	# Reduced / Extracted Features			
	25	50	100	200
Avg. Accuracy (%)	75.01	73.97	74.32	74.92
Min. Accuracy (%)	59.52	57.14	57.14	60.71
Max. Accuracy (%)	89.29	92.58	89.28	88.09
Variance of Accuracy	0.00247	0.00221	0.00187	0.00244

TABLE4. TESTING ACCURACY OF THE AUTOENCODER FEATURE EXTRACTION METHOD WITH 25-FEATURES, 50-FEATURES, 100-FEATURES, 200-FEATURES

	# Reduced / Extracted Features			
	25	50	100	200
Avg. Accuracy (%)	75.84	76.19	76.16	76.68
Min. Accuracy (%)	57.14	55.35	50.00	53.57
Max. Accuracy (%)	90.47	87.5	91.07	92.85
Variance of Accuracy	0.00273	0.00263	0.00254	0.00359

TABLE5. TESTING ACCURACY OF THE ICA FEATURE REDUCTION METHOD WITH 25-FEATURES, 50-FEATURES, 100-FEATURES, 200-FEATURES

	# Reduced / Extracted Features			
	25	50	100	200
Avg. Accuracy (%)	74.44	73.21	72.59	73.69
Min. Accuracy (%)	50.00	50.00	51.78	52.31
Max. Accuracy (%)	90.47	92.75	89.28	89.28
Variance of Accuracy	0.00327	0.00309	0.00299	0.00333

b. Results of Different Numbers of Extracted Features

For 8 out of 15 participants, the 100-extracted features option yielded the highest personal average prediction accuracy, while 3 participants achieved their highest accuracy with the 50-extracted features option, 2 participants achieved highest accuracy with 25-extracted features option, and another 2 participants achieved highest accuracy with the 200-extracted features option. For all the options on the number of extracted features, the numbers of participants achieved 75% accuracy or higher were nine persons for 25-features, eight persons for 50-features, twelve persons for 100-features, and ten persons for 200-features.

Among all the options of feature reduction/extraction numbers tested in the study, the 200-extracted features option yielded the highest average accuracy – 76.68.

It is easy to notice that different selections on the numbers of extracted features do not make a huge impact on the final performance of the algorithm. The reason behind this is probably that although many features have been extracted, those key features which actually play important roles in the prediction process only take up a small portion within the entire feature group. This is great news to us since the majority of the running time of the algorithm falls in the feature extraction step. Thus, in the future, the execution time of the algorithm can be greatly reduced by shrinking the size of the extracted feature array.

c. Summary of the Overall Performance

The personal best average accuracies of comfort level prediction for all fifteen participants were above 80% considering all hyperparameter combinations. There are 10 out of 15 participants whose best average accuracies are above 78%. The best personal average accuracy of comfort level prediction among all fifteen participants and hyperparameter combinations was 92.86%, achieved from Participant 1 when the Autoencoder method and 200 extracted features option were applied.

In this study, the overall comfort level prediction accuracies achieved with SVMRFE, Autoencoder, ICA feature extraction methods and SVM multi-class classifier were 75.01%, 76.68% and 74.44% respectively as shown in Table 3-5. AE-based method not only yielded the best overall performance among all participants, but also provided the best average accuracies in 10 out of 15 participants. In addition, the AE-based method yielded the highest maximum accuracy across all participants. Based on these results, we can conclude that AE-based method is superior with certain participants in its performance, but lacks stability compared to the SVMRFE method due to the fact that its minimum accuracies are lower than the SVMRFE method. The reason AE method yields higher overall prediction accuracy could be due to its powerful feature compression capability. While SVM-based feature reduction method only reduces the size of the feature set by simply eliminating less useful features, the AE method compresses and maps the original feature set to a smaller space. And with such a smaller set, the classification algorithm can achieve a higher performance. This study proves the potential of the Autoencoder model for future wearable sensing studies. According to the statistics results in Table 3-5, the performance ranking from high to low is AE-based method, SVMRFE, and then ICA. In terms of choices on feature reduction numbers, the results demonstrated that the influence of this number is much less critical than the choice on different feature extraction methods. We also compared the accuracy results of the 58 scenarios with three different methods and four feature number setups, as shown in Fig. 8-10. The performance gaps of different methods and feature number setups are negligible compared to the gaps among different scenarios. It is easy to notice that the accuracy curves drop into a valley region between scenarios #15-16 and #20-28, which correspond to the delivery distance factor group. These results demonstrated that the difficulty of predicting distance-related comfort is much higher than other factors.

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2024.3383296

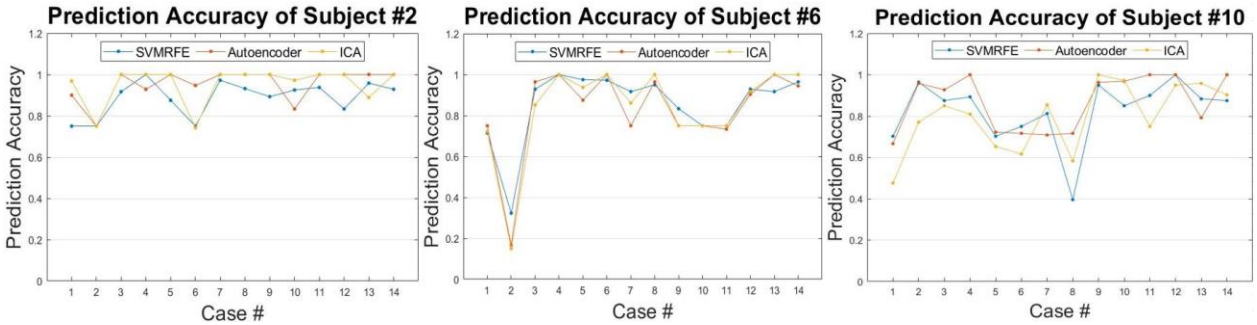

Fig. 7. Prediction Results of Test Subject #2, 6 & 10 in Single Test Run.

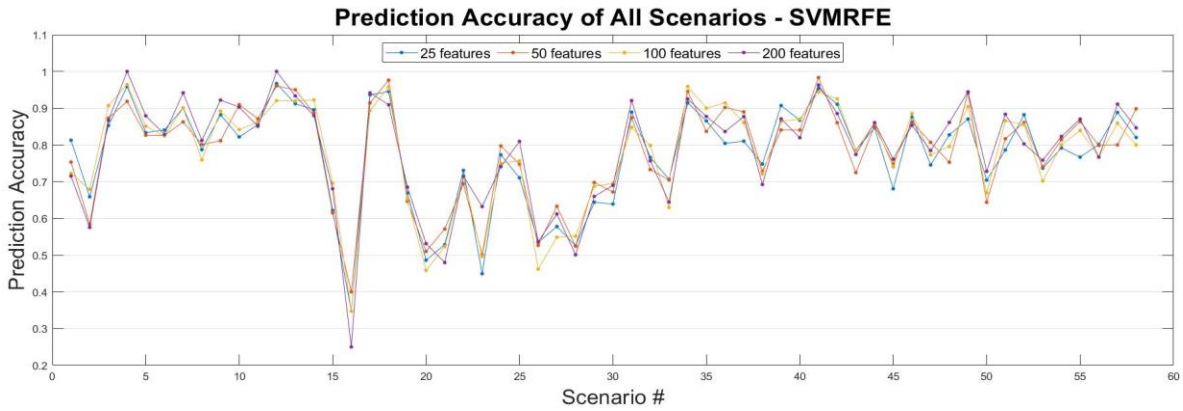

Fig. 8. Prediction Results of 58 Scenarios with SVMRFE Method

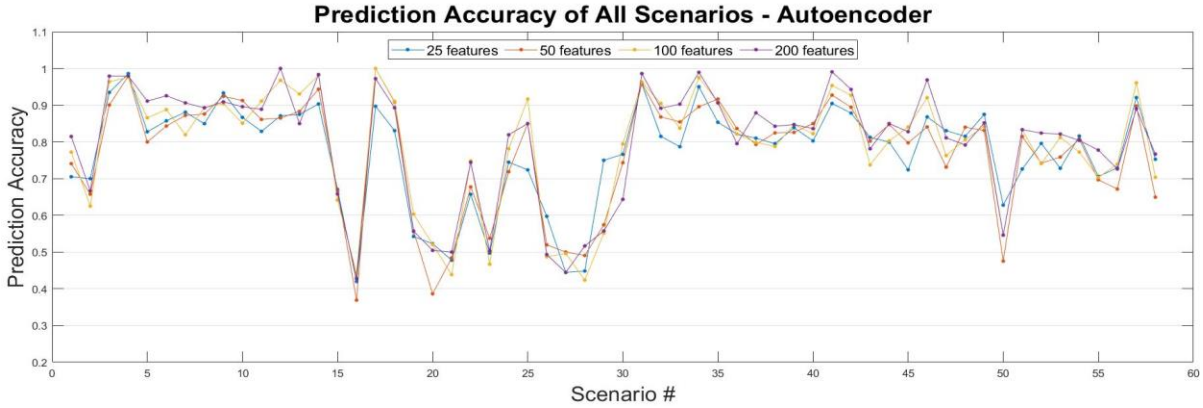

Fig. 9. Prediction Results of 58 Scenarios with Autoencoder Method

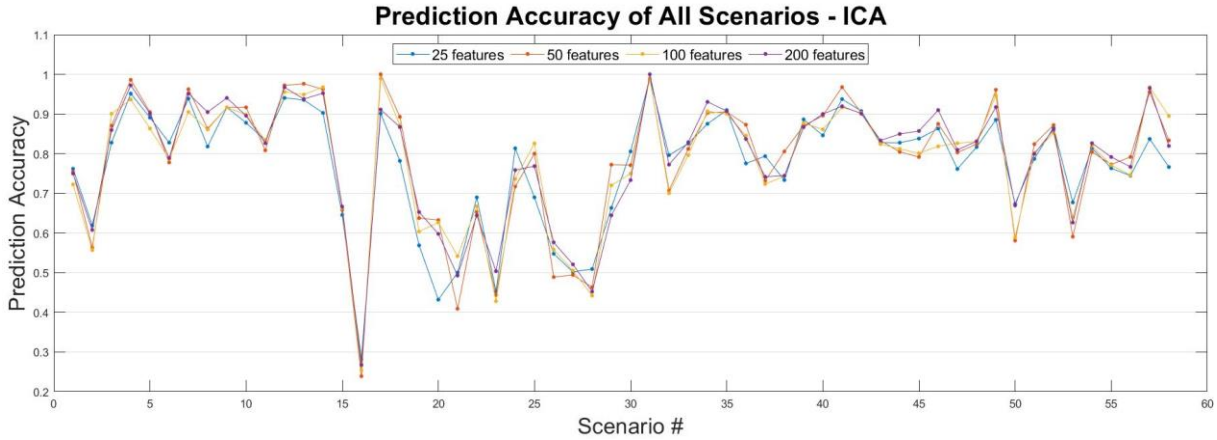

Fig. 10. Prediction Results of 58 Scenarios with ICA Method

There are three participants whose subjective ratings are extremely unbalanced among the range [1,5], one of them

mostly falls in the range between [2-4], the other two were too biased to one side of the range. Such an unbalanced dataset will negatively impact the result performance. More pre-test training will be provided to the participants in the future to guarantee the quality of the collected data. The good side is that all subjects provided overall comfort ratings within the range from 2 to 4 for the two reference cases, which means no one has overly concentrated and biased ratings near the two borders of the range.

The approach we proposed in this study demonstrated its value in competitive performance, which also focuses on human emotion prediction. Furthermore, considering the fact that human comfort prediction, which involves a variety of factors and states, is a much more complicated problem than emotion detection, the results achieved in our study validate the feasibility and effectiveness of the proposed approach in using physiological signals to detect the comfort levels of humans during physical human-robot collaborations.

Considering that physiological signals are susceptible to noises and uncertainties which could be affected by many unknown factors and random events such as body movements, environment noises, etc., the physiological-based approach can be combined with an analytical model to make corrections to the results from the physiological-data prediction model in order to reduce the negative effects from these uncertainties.

V. Conclusion

In this paper, a physiological-data-based general human comfort prediction model is proposed under human-robot collaboration scenarios. Previous related studies mostly utilize subjective ratings method to evaluate how human comfort varies as one robot factor changes, yet such methods are limited in evaluating comfort online. The proposed method in this paper tackled these two limitations at the same time, measuring and evaluating human comfort under the effects of multiple factors online.

In this study, an ABB Yumi robot is used as the collaborative robot for the HRC tasks. A sequence of 58 robot delivery tasks were designed with five varying robot factors. Wearable sensing system was used in the experiment to collect physiological data and the subjective comfort level of the participants was collected with self-reporting forms. Additionally, we developed an SVM-based machine-learning model to predict general human comfort levels based on the data acquired during the experiment. Based on the prediction accuracy results, the method proposed in this study was proved to be effective in HRC scenarios. The overall comfort level prediction accuracies achieved with SVMRFE, Autoencoder, ICA feature extraction methods and SVM multi-class classifier were 75.01%, 76.68% and 74.44% respectively, thus proving that SVMRFE, Autoencoder-based and ICA feature reduction/extraction methods are effective in bio-signal applications. Among the three methods, the AE-based method yields the highest prediction accuracy.

In the future, we aim to further improve the accuracy of prediction by combining the analytical prediction model with physiological model to refine the results from physiological-data-based model, or by using new physiological feature extraction methods to improve current framework's accuracy. In addition, force/haptic/tactile interactions with different parameter settings could be designed and the outcome in this paper could be used to assess human comfort levels from a physiological perspective while humans are conducting such tasks. In future works, we will also try to reduce the feature processing lag of EEG and test its effectiveness in real-time applications.

VI. Acknowledgment

This work was supported by the National Science Foundation under Grant IIS-1845779.

References

[1] ISO, "ISO 10218-1:2011 Robots and robotic devices." Jul. 2011, [Online]. Available: https://www.iso.org/standard/51330.html.

[2] Wang, Haiying, et al. "Experimental Comparison of Local Direct Heating to Improve Thermal Comfort of Workers." *Building and Environment*, vol. 177, Elsevier Ltd, 2020, p. 106884–, doi:10.1016/j.buildenv.2020.106884.

[3] Lan, Li, Pawel Wargocki, and Zhiwei Lian. "Optimal thermal environment improves performance of office work." *Rehva Journal* 49.1 (2012): 12-17.

[4] Ye, Xiaojiang, et al. "Thermal Environment and Productivity in the Factory." *ASHRAE Transactions*, vol. 116, no. 1, American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Inc, 2010, p. 590–.

[5] H. Bellem, M. Kl¨uver, M. Schrauf, H.-P. Sch¨oner, H. Hecht, and J. F. Krems, "Can we study autonomous driving comfort in moving-base driving simulators? a validation study," Human factors, vol. 59, no. 3, pp. 442–456, 2017.

[6] M. P. De Looze, L. F. Kuijt-Evers, and J. Van Dieen, "Sitting comfort and discomfort and the relationships with objective measures," Ergonomics, vol. 46, no. 10, pp. 985–997, 2003.

[7] R. R. Bishu, M. S. Hallbeck, M. W. Riley, and T. L. Stentz, "Seating comfort and its relationship to spinal profile: A pilot study," International Journal of Industrial Ergonomics, vol. 8, no. 1, pp. 89–101, 1991.

[8] D. Oborne, "Vibration and passenger comfort," Applied Ergonomics, vol. 8, no. 2, pp. 97–101, 1977.

[9] Wang,W.; Liu, N.; Li, R.; Chen, Y.; Jia, Y, "Hucom: A model for human comfort estimation in human-robot collaboration." *In Proceedings of the 2018 Dynamic Systems and Control (DSC) Conference, Atlanta, GA,USA,*, 30 September–3 October 2018.

[10] Mead, Ross, and Maja J. Mataric. "Proxemics and Performance: Subjective Human Evaluations of Autonomous Sociable Robot Distance and Social Signal Understanding." *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* IEEE, 2015, pp. 5984–91, doi:10.1109/IROS.2015.7354229.

[11] Stark, Jessi, et al. "Personal Space Intrusion in Human-Robot Collaboration." *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction,* ACM, 2018, pp. 245–46, doi:10.1145/3173386.3176998.

[12] Lasota, Przemyslaw A., and Julie A. Shah. "Analyzing the Effects of Human-Aware Motion Planning on Close-Proximity Human–Robot Collaboration." *Human Factors,* vol. 57, no. 1, SAGE Publications, 2015, pp. 21–33, doi:10.1177/0018720814565188.

[13] Alami, Rachid, et al. "Task planning for human-robot interaction." Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies. 2005.

[14] Ciccarelli, Marianna, et al. "A system to improve the physical ergonomics in Human-Robot Collaboration." Procedia Computer Science 200 (2022): 689-698.

[15] C. L. Lisetti and F. Nasoz, "Using non-invasive wearable computers to recognize human emotions from physiological signals," EURASIP Journal on Advances in Signal Processing, vol. 2004, no. 11, p. 929414, 2004.

[16] J. Kim and E. Andr´e, “Emotion recognition based on physiological changes in music listening,” IEEE transactions on pattern analysis and machine intelligence, vol. 30, no. 12, pp. 2067–2083, 2008.

[17] Shan, Xin, and En-Hua Yang. “Supervised Machine Learning of Thermal Comfort Under Different Indoor Temperatures Using EEG Measurements.” Energy and buildings 225 (2020): 110305–. Web.

[18] C. Maaoui and A. Pruski, “Emotion recognition through physiological signals for human-machine communication,” Cutting Edge Robotics, vol. 2010, pp. 317–332, 2010

[19] Kang, Min-Koo, et al. "A wellness platform for stereoscopic 3D video systems using EEG-based visual discomfort evaluation technology." Applied ergonomics 62 (2017): 158-167.

[20] Wikipedia contributors. “Robot Operating System.” *Wikipedia*, 1 May 2021, en.wikipedia.org/wiki/Robot_Operating_System.

[21] “ABB's Collaborative Robot -YuMi.” *Robotics,* new.abb.com/products/robotics/collaborative-robots/irb-14000-yumi.

[22] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, “Eeg-based emotion recognition in music listening,” IEEE Transactions on Biomedical Engineering, vol. 57, no. 7, pp. 1798–1806, 2010.

[23] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tr¨oster, and U. Ehlert, “Discriminating stress from cognitive load using a wearable eda device,” IEEE Transactions on information technology in biomedicine, vol. 14, no. 2, pp. 410–417, 2009.

[24] E.-H. Jang, B.-J. Park, M.-S. Park, S.-H. Kim, and J.-H. Sohn, “Analysis of physiological signals for recognition of boredom, pain, and surprise emotions,” Journal of physiological anthropology, vol. 34, no. 1, p. 25, 2015.

[25] M. Benedek and C. Kaernbach, “A continuous measure of phasic electrodermal activity,” Journal of neuroscience methods, vol. 190, no. 1, pp. 80–91, 2010.

[26] M. Ali, F. Al Machot, A. Mosa, and K. Kyamakya, CNN Based Subject- Independent Driver Emotion Recognition System Involving Physiological Signals for ADAS, 01 2016, pp. 125–138.

[27] B. Figner, R. O. Murphy et al., “Using skin conductance in judgment and decision making research,” A handbook of process tracing methods for decision research, pp. 163–184, 2011.

[28] “ClassificationECOC.” *Multi-class Model for Support Vector Machines (SVMs) and Other Classifiers - MATLAB*, https://www.mathworks.com/help/stats/classificationecoc.html.

[29] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recognition Letters*, Vol. 30, Issue 3, 2009, pp. 285–297.

[30] Guyon, Isabelle, et al. “Gene Selection for Cancer Classification Using Support Vector Machines.” Machine Learning, vol. 46, no. 1, Kluwer Academic Publishers, 2002, pp. 389–422, https://doi.org/10.1023/A:1012487302797.

[31] R. W. Picard, E. Vyzas, and J. Healey, “Toward machine emotional intelligence: Analysis of affective physiological state,” IEEE transactions on pattern analysis and machine intelligence, vol. 23, no. 10, pp. 1175– 1191, 2001.

[32] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

[33] Yao, Shengnan, et al. “Validating the Performance of One-Time Decomposition for fMRI Analysis Using ICA with Automatic Target Generation Process.” Magnetic Resonance Imaging, vol. 31, no. 6, Elsevier Inc, 2013, pp. 970–75, https://doi.org/10.1016/j.mri.2013.03.014.

[34] H. Su and Y. Jia, "Study of Human Comfort in Autonomous Vehicles Using Wearable Sensors," in IEEE Transactions on Intelligent Transportation Systems, doi: 10.1109/TITS.2021.3104827.