Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 8371637, 12 pages
https://doi.org/10.1155/2021/8371637

Research Article

Comparison Analysis of Different Time-Scale Heart Rate Variability Signals for Mental Workload Assessment in Human-Robot Interaction

Shiliang Shao[1,2], **Ting Wang**[1,2], **Yawei Li**[1,2,3], **Chunhe Song**[1,2], **Yihan Jiang**[1,2,4] and **Chen Yao**[1,2]

[1]*State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China*
[2]*Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China*
[3]*University of Chinese Academy of Sciences, Beijing 100049, China*
[4]*Shenyang Ligong University, School of Automation and Electrical Engineering, Shenyang 110159, China*

Correspondence should be addressed to Shiliang Shao; shaoshiliang@sia.cn and Ting Wang; wangting@sia.cn

Received 21 June 2021; Revised 19 August 2021; Accepted 31 August 2021; Published 6 October 2021

Academic Editor: Fa Zhu

Excessive mental workload affects human health and may lead to accidents. This study is motivated by the need to assess mental workload in the process of human-robot interaction, in particular, when the robot performs a dangerous task. In this study, the use of heart rate variability (HRV) signals with different time scales in mental workload assessment was analyzed. A humanoid dual-arm robot that can perform dangerous work was used as a human-robot interaction object. Electrocardiogram (ECG) signals of six subjects were collected in two states: during the task and in a relaxed state. Multiple time-scale (1, 3, and 5 min) HRV signals were extracted from ECG signals. Then, we extracted the same linear and nonlinear features from the HRV signals at different time scales. The performance of machine learning algorithms using the different time-scale HRV signals obtained during the human-robot interaction was evaluated. The results show that for the per-subject case with a 3 min HRV signal length, the K-nearest neighbor classifier achieved the best mental workload classification performance. For the cross-subject case with a 5 min time-scale signal length, the gentle boost classifier achieved the best mental workload classification accuracy. This study provides a novel research idea for using HRV signals to measure mental workload during human-robot interaction.

1. Introduction

Nowadays, robots, instead of humans, work in unstructured environments, expanding the scope of human work. Humans interact with robots through visual, tactile, and other feedback [1–4]. The robot can be operated remotely to complete a dangerous task; this operation can be challenging for humans. At present, research in the field of robotics primarily focuses on how robots perform human control instructions, how they perceive environmental information, and how autonomous operation can be achieved [5, 6]. However, this research neglects the robot's assessment of the human's psychological activity and the emotions of humans interacting with the robot. Therefore, it is of great significance to accurately measure the mental workload of the operator during their interaction with the robot [7, 8].

Mental workload can be measured continuously and objectively using physiological signals. In particular, heart rate variability (HRV) signals have been widely studied because they are easy to collect. In [9], the relationships between mental workload and time-domain, frequency-domain, and Poincare plot features of 5 min signals were analyzed. In [10], 5 min HRV signal segments were used to detect the mental workload of a worker. Several linear features (time and frequency domains) were utilized. Then, the combination of principal component analysis and support vector machine (SVM) achieved 84.4% accuracy. In fact, the physiological system of the human body can be

regarded as a nonlinear system. However, the nonlinear nature of HRV signals cannot be reflected by linear analysis methods [11, 12]. In [13], the mental workload of performing MATA-II tasks was measured using 5 min scale HRV signals. This study extracted the multiscale entropy features of the HRV. Using those, it obtained a higher accuracy for mental workload recognition than using traditional time- and frequency-domain features. In [14], 5 min length HRV signal segments were utilized to evaluate the mental workload of hospital staff. A variety of conventional and multiscale HRV features were extracted, and SVM was used as the classifier. The results showed that the multiscale features obtain a better mental workload recognition effect. In [15], the respiratory and HRV signals were extracted using 5 min scale electrocardiogram (ECG) signals. This study introduced a novel method that fused respiratory and HRV signals to assess subtle variations in sympathovagal balance using ECG recordings during the MATA-II mission. Standard short-term HRV analysis is usually performed on 5 min recordings [16], and shorter recordings of HRV signals are being researched, aiming at a faster detection of mental workload. In [17], human HRV signals were collected during human-robot interaction through different types of wearable devices. Using signals of 3 min length, the linear features of HRV signals collected by different wearable devices were extracted, compared, and analyzed under different mental workload levels. In [18], 3 min HRV signals were used, and linear and nonlinear features were utilized. Several machine learning algorithms have been utilized for assessing the mental workload of humans while operating a dual-arm robot. In [19], 2.5 min HRV signals were detected by a consumer smart watch. Subsequently, the mental workload of human interaction with multiple robots was studied. However, analysis of mental workload recognition with HRV signals at different time scales is not sufficiently researched. In [20], a nonparametric statistical test method was utilized to analyze the significant differences between rest and stress phases with time scales of 30 s and 1, 2, 3, and 5 min. However, HRV signals were obtained from healthy subjects during an examination and in a resting condition, not during human-robot interaction.

Humans use visual, haptic, and other feedback information to remotely perceive the environment information during human-robot interaction, and the robot is remotely operated to complete the task. The entire human-robot interaction process requires the joint perception and decision-making of human hands, eyes, ears, brain, and other limbs and organs, which may be very challenging for the operator. At present, there is a lack of mental workload measurement analysis during human-robot interactions using HRV with different time scales. Therefore, in this study, the differences among HRV signals of multiple time scales in measuring mental workload were analyzed; six traditional machine learning methods were used to evaluate the performances of HRV signals with different time scales. Traditional machine learning methods were used because they are more suitable for small sample sizes. Although deep learning methods have been widely studied, many training samples are required.

The contribution of this study can be summarized as follows:

(i) During human-robot interaction, HRV signals were collected based on a single physiological signal. In addition, linear and nonlinear features of different time-scale HRV signals were extracted, and statistical differences between the mental workloads in the two states were analyzed

(ii) A variety of representative machine learning algorithms were applied. Differences in the performances of the machine learning algorithms with statistically different linear and nonlinear features extracted from HRV signals of different time scales in evaluating mental workload were analyzed

(iii) Finally, the different performances of the algorithms with HRV signals of various time scales in evaluating mental workload are discussed

The remainder of this paper is organized as follows. Section 2 introduces the data collection and preprocessing algorithms. The mental workload assessment results of algorithms using different time scales of HRV signals and a discussion of the results are presented in Section 3. The concluding remarks are presented in Section 4.

2. Data and Method

The research block diagram is shown in Figure 1. It can be seen that the ECG signals were obtained from volunteers while they operated the dual-arm robot and in the rest state. HRV signals were then extracted from the ECG signals. Using a sliding window of different time scales (1, 3, and 5 min), the HRV signals were divided to obtain a collection of sample data of different time scales. Then, linear and nonlinear features of different time scales were extracted. In addition, an SVM, K-nearest neighbor (KNN) classifier, gentle boost (GB), linear discriminant analysis (LDA), naive Bayes (NB), and decision tree (DT) were utilized to identify the task-performing and rest states. The performance differences of the classifiers in the mental workload evaluation with HRV signals at different time scales were compared and analyzed.

2.1. Data. In this subsection, the subjects and data acquisition processes are described. Then, a preprocessing algorithm is introduced to obtain the HRV signals from the collected ECG signals. In addition, multiple time-scale HRV signal segments are obtained using sliding windows of different time scales.

2.1.1. Participants. The ECG signals used for mental workload assessments were obtained from six male participants. A description of the six subjects is provided in Table 1. They were recruited from the Shenyang Institute of Automation, Chinese Academy of Sciences. Their average age was 25.16 (±2.93). They had normal or corrected vision and were all healthy, with no nervous system diseases. Before starting the experimental data collection, all participants were

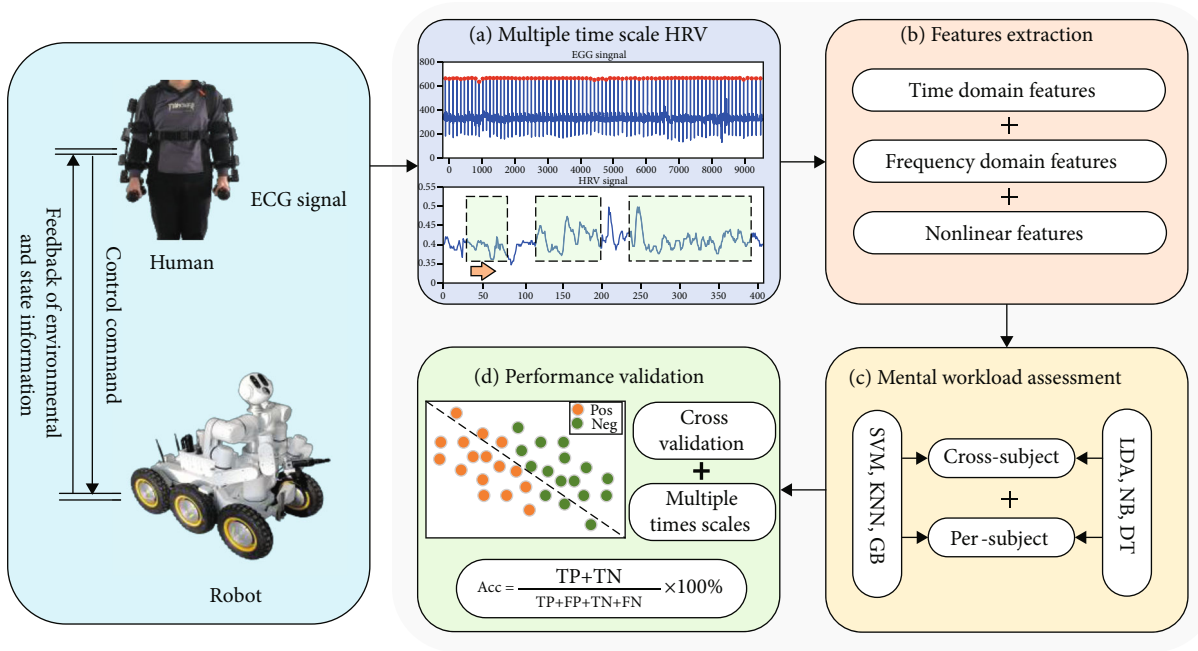

Figure 1: Framework of multiple time-scale HRV analysis for mental workload assessment. HRV: heart rate variability; ECG: electrocardiogram; SVM: support vector machine; KNN: K-nearest neighbors; GB: gentle boost; LDA: linear discriminant analysis; NB: naive Bayes; DT: decision tree; TP: true positive; TN: true negative; FP: false positive; FN: false negative.

Table 1: Participant characteristics.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6
Height (cm)	180	175	173	180	175	178
Weight (kg)	67.5	78.5	58	55	75	72.5
Age (years)	24	24	31	23	24	25
Body mass index (kg/m^2)	20.8	25.6	19.4	17.0	24.5	22.9

informed of the entire data collection process and precautions.

2.1.2. Data Acquisition. In this study, the operating object was a dual-arm robot shown in Figure 2. It can be seen that the robot consists of six wheels and two arms. Moreover, each wheel is independent, and each arm has seven degrees of freedom to access all positions in space. In addition, the top of the robot is equipped with a binocular camera for environmental observations. The robot controller is an exoskeleton device that can be worn by an operator (Figure 3). The exoskeleton controller also has two arms, and each arm has seven degrees of freedom, similar to the dual-arm robot. The ECG signal collection process is shown in Figure 4. A portable sensor was placed on the chest of the operator for the acquisition of ECG signals. The captured ECG signals were sent to a computer via Bluetooth for processing. ECG signals were collected in two states of the operator: during the operation of the dual-arm robot and during rest.

2.1.3. Signal Preprocessing. The HRV signals refer to a time series consisting of intervals between each pair of heartbeats. Therefore, to obtain the HRV signals, it is necessary to detect the peak and trough values of the ECG signals. Therefore, the Q, R, and S waves of the ECG signal were detected using a QRS wave group detection method [21]. However, there may be an abnormal point in the output RR interval sequence. Therefore, a classical median-filtering algorithm was applied to the output RR interval sequence [22]. The RR interval sequence was regarded as an HRV signal. As shown in Figures 5–7, sliding windows at different time scales (1, 3, and 5 min) were used with an overlap of 30 s. HRV signals were then divided into six groups: M-1, R-1, M-3, R-3, M-5, and R-5 groups. The M group signals represent the operator in the task-performing state, and the R group signals represent the operator in the rest state.

The proposed mental workload assessment preprocessing algorithm is described in Algorithm 1, where $x_i(t)$ is the ECG data recorded from the ith participant, and I is the number of participants. The purpose of Steps 1 to 6 is to obtain the HRV signals $y_i(t)$ from $x_i(t)$ signals. The HRV signals $y_i(t)$ are segmented into different time-scale (1, 3, and 5 min) segments $y_i^1(t)$, $y_i^3(t)$, and $y_i^5(t)$ in Steps 7 to 10.

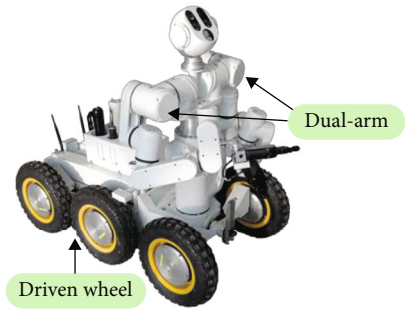

Figure 2: Dual-arm robot experiment platform.

Figure 3: Exoskeleton robot controller.

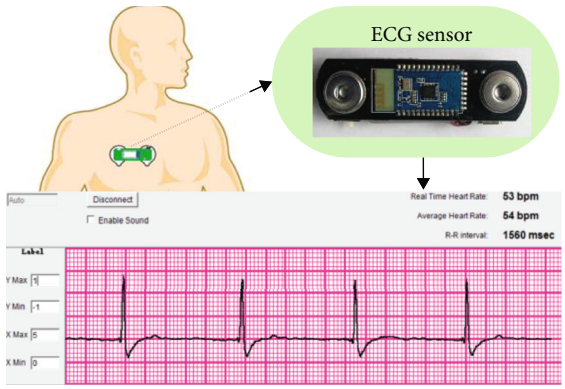

Figure 4: Process of ECG signal collection. ECG: electrocardiogram.

2.2. Method. Linear and nonlinear analysis methods are the most commonly used HRV signal analysis methods. Therefore, in this subsection, the linear and nonlinear features used in this study are described. The collection of physiological signals during human-robot interaction requires considerable manpower and energy; thus, it is difficult to collect large-scale sample data. However, machine learning algorithms do not require large-scale sample data for efficient feature recognition [23, 24]. Therefore, in this study, several different types of machine learning algorithms (SVM, KNN, GB, LDA, NB, and DT) were used to compare the effects of feature recognition.

2.2.1. Feature Extraction. First, the linear and nonlinear features used in this study are presented. In human-robot interaction, the fluctuation of the operator's mental workload is related to the fluctuation of the human autonomic nervous system (ANS). The ANS consists of the sympathetic and parasympathetic nervous systems. The time- and frequency-domain features of HRV signals can reflect fluctuations in the sympathetic and the parasympathetic nervous systems. In addition, nonlinear features can reflect the nonlinear dynamic characteristics of the HRV signal [25, 26].

The linear features include time- and frequency-domain features. First, we introduce time features.

SDNN denotes the standard deviation of all RR intervals:

$$\mathrm{SDNN} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{RRs}_i - \frac{1}{N}\sum_{i=1}^{N}\mathrm{RRs}_i\right)}. \tag{1}$$

RMSSD denotes the root mean square of the adjacent RR interval difference:

$$\mathrm{RMSSD} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}\left(\mathrm{RRs}_{i+1} - \mathrm{RRs}_i\right)^2}. \tag{2}$$

pNN50 denotes the ratio of the number of pairs of adjacent RR intervals with a difference of more than 50 ms:

$$\mathrm{PNN50} = \frac{\mathrm{num}[(\mathrm{RRs}_{i+1} - \mathrm{RRs}_i) > 50\,\mathrm{ms}]}{N-1}. \tag{3}$$

In addition, all RR intervals were integrated and divided by the maximum density distribution parameter, and the mean and median of the HRV signals were also extracted as time-domain features.

In this study, all frequency-domain features were obtained based on the power spectral density [27]. Furthermore, the basic frequency-domain features are defined as the sum of the power spectra at different frequency ranges: aTotal = 0 – 0.4 Hz; aVLF = 0.003 – 0.04 Hz; aLF = 0.04 – 0.15 Hz; and aHF = 0.15 – 0.4 Hz. The ratio of aLF and aHF is defined as

$$\frac{\mathrm{LF}}{\mathrm{HF}} = \frac{\mathrm{aLF}}{\mathrm{aHF}}. \tag{4}$$

The percentage of aVLF, aLF, and aHF are defined as

$$\begin{aligned} \mathrm{pVLF} &= \frac{\mathrm{aVLF}}{\mathrm{aTotal}}, \\ \mathrm{pLF} &= \frac{\mathrm{aLF}}{\mathrm{aTotal}}, \\ \mathrm{pHF} &= \frac{\mathrm{aHF}}{\mathrm{aTotal}}. \end{aligned} \tag{5}$$

The respective ratios of aLF and aHF to aLF + aHF are

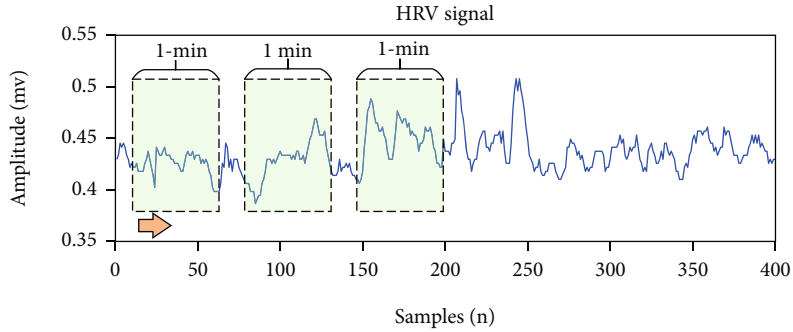

Figure 5: Heart rate variability segment by 1 min time scale.

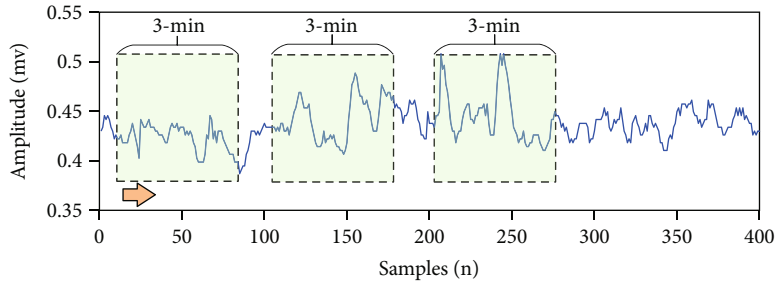

Figure 6: Heart rate variability segment by 3 min time scale.

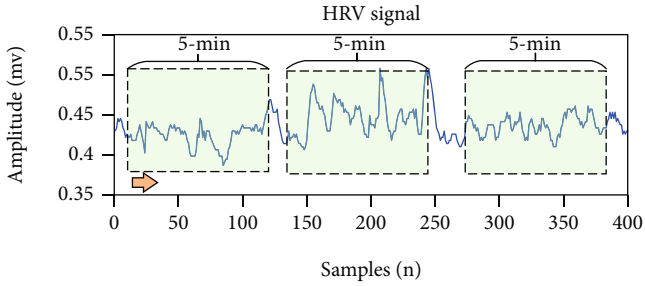

Figure 7: Heart rate variability segment by 5 min time scale.

defined as

$$\begin{aligned} \text{nLF} &= \frac{\text{aLF}}{\text{aLF} + \text{aHF}}, \\ \text{nHF} &= \frac{\text{aHF}}{\text{aLF} + \text{aHF}}. \end{aligned} \tag{6}$$

Finally, two typical nonlinear analysis methods applied in this study are presented. These are sample entropy (SaEn) and detrended fluctuation analysis (DFA). On the one hand, SaEn is a method for investigating the dynamics of HRV signals. It has the advantages of strong antinoise and antijamming abilities. In addition, it can be used to analyze shorter HRV signals. In the case of large differences in the parameter value range, good consistency is still achieved. On the other hand, DFA is suitable for the analysis of nonstationary time series, and HRV signals have this characteristic. In addition, the DFA method can filter out the trend components in the HRV signal. Therefore, it can effectively avoid the disturbance of false correlations owing to noise and signal instability.

2.2.2. Mental Workload Recognition. In this subsection, the abstracted feature vector of HRV signals at different time scales is used to evaluate the mental workload. The different time-scale HRV features were analyzed using the t-test to obtain the statistical significance of the difference between task-performing and relaxed states; $p < 0.05$ was considered statistically significant [28]. Then, linear and nonlinear features with statistical differences were used to construct feature vectors as inputs to machine learning algorithms. Six different machine learning methods, SVM, KNN, LDA, GB, NB, and DT, were used in this study to exclude the effects of performance differences in machine learning algorithms.

After the initial HRV signal preprocessing, 1, 3, and 5 min time-scale HRV signals for mental workload can be assessed using Algorithm 2. The linear and nonlinear features of the ith subject were extracted in Steps 2 to 5. $\tilde{\mathbf{F}}_i^s$ is defined as the feature vector in the human task-performing state, and $\mathbf{F}_i^s$ is defined as the feature vector in the human relaxed state. Steps 6 to 11 define the process of per-subject mental workload assessment. $\mathbf{F}_{i\_\text{Train}}^s$ and $\mathbf{F}_{i\_\text{Test}}^s$ are the training and testing sets of the ith subject, respectively. Steps 12 to 18 define the process of cross-subject mental workload assessment. The extracted HRV features $\tilde{\mathbf{F}}_i^s$ and $\mathbf{F}_i^s$ of all subjects in task-performing and relaxation states are merged

*Input:*ECG signals $x_i(t)$ for each i subject.
*Output:*Multiple time-scale HRV signals for all subjects.
1: *For* each i such that $1 \le i \le I$ *do*
2: Q wave, R wave, and S wave of ECG signals $x_i(t)$ are detected.
3: RR internal sequence is obtained.
4: Abnormal points in the output RR internal sequence are removed by median filtering.
5: HRV signals $y_i(t)$ are obtained.
6: *End for*
7: *For* each i such that $1 \le i \le I$ *do*
8: Segment the $y_i(t)$ signals into 1 min, 3 min, and 5 min time-scale segments defined as
9: $y_i^1(t)$, $y_i^3(t)$, and $y_i^5(t)$, respectively.
10: *End for*

Algorithm 1: Mental workload assessment preparation.

*Input:*Multiple time-scale HRV segments for all subjects.
*Output:*Per-subject and cross-subject probability of mental workload.
1: *For* each time scale s, $s = 1, 3, 5$.
2: *For* each i such that $1 \le i \le I$ *do*
3: Extract linear and nonlinear features $\tilde{\mathbf{F}}_i^s$ for each $y_i^s(t)$signal at task-performing state.
4: Extract linear and nonlinear features $\mathbf{F}_i^s$ for each $y_i^s(t)$signal at relaxation state.
5: *End for*
6: *If* per-subject mental workload assessment
7: *For* each i such that $1 < i < I$ *do*
8: Train classifiers (SVM, KNN, LDA, GB, NB, and DT) based on the training set $\mathbf{F}_{i\_\text{Train}}^s$ randomly selected from $\mathbf{F}_i^s$.
9: Obtain the probability of mental workload based on the testing set $\mathbf{F}_{i\_\text{Test}}^s$, which is defined as $\mathbf{F}_i^s$-$\mathbf{F}_{i\_\text{Train}}^s$.
10: *End for*
11: *End if*
12: *If* cross-subject mental workload assessment
13: Merge matrices $\tilde{\mathbf{F}}_1^s$, $\tilde{\mathbf{F}}_i^s$,…, $\tilde{\mathbf{F}}_I^s$ into one matrix $\tilde{\mathbf{F}}^s$.
14: Merge matrices $\mathbf{F}_1^s$, $\mathbf{F}_i^s$,…, $\mathbf{F}_I^s$ into one matrix $\mathbf{F}\wedge^s$.
15: Train machine learning method (SVM, KNN, LDA, GB, NB, and DT) based on the training set $\mathbf{F}_{\text{Train}}^s$ and $\tilde{\mathbf{F}}_{\text{Train}}^s$ randomly selected from $\mathbf{F}\wedge^s$ and $\tilde{\mathbf{F}}^s$, respectively.
16: Obtain probability of mental workload based on the testing set $\mathbf{F}_{\text{Test}}^s$ and $\tilde{\mathbf{F}}_{\text{Test}}^s$, which are defined as $\mathbf{F}\wedge^s$-$\mathbf{F}_{\text{Train}}^s$ and $\tilde{\mathbf{F}}^s$-$\tilde{\mathbf{F}}_{\text{Train}}^s$, respectively.
17: *End if*
18: *End for*

Algorithm 2: Mental workload assessment after preprocessing.

into matrices $\tilde{\mathbf{F}}^s$ and $\mathbf{F}\wedge^s$ in Steps 13 and 14, respectively. Then, in Steps 15 to 17, the merged matrices $\tilde{\mathbf{F}}^s$ and $\mathbf{F}\wedge^s$ are prepared for model construction and mental workload assessment.

3. Experimental Results

In this section, the mental workload recognition performance of classifiers with HRV signals of different time scales is presented. The statistical differences of the linear and nonlinear features extracted in this study among different mental workload levels were analyzed via a t-test, and the feature vectors were composed of per-subject and cross-subject mental workload assessments.

To evaluate the performance of mental workload classification with different time scales, accuracy was used, which is defined as follows:

$$\text{Accuracy}: \text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100\%, \quad (7)$$

where TP is true positive, FP is false positive, FN is false negative, and TN is true negative.

3.1. Per-Subject Mental Workload Evaluation. The results of per-subject mental workload evaluation at different time scales (1, 3, and 5 min) are presented. The samples of each subject were randomly divided into two sets. One was used for training the machine learning model, and the other was

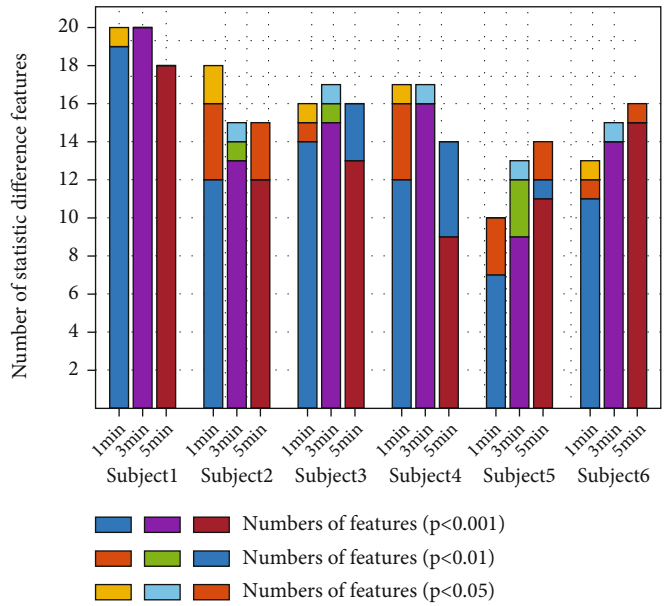

Figure 8: Results with statistically significant features of per-subject analysis at different time scales.

used for testing the model. In addition, to increase the reliability of the results, the average of the results repeated 500 times was regarded as the final classification result.

3.1.1. Results of Statistically Significant Features of 1, 3, and 5 min Length. Figure 8 shows the statistics of the significantly different ($p < 0.001$, $p < 0.01$, and $p < 0.05$) features at different time scales of each subject. It can be seen that Subject 1 has more significantly different ($p < 0.001$) features at the 3 min time scale, followed by the 1 min and 5 min time scales. Subject 2 showed more significantly different features at the 3 min time scale and at the 1 min time scale; the sum of the most significantly different ($p < 0.001$) features and the significantly different ($p < 0.01$ and $p < 0.05$) features was the largest. Subject 3 and Subject 4 have the most significantly different ($p < 0.001$) features at the 3 min time scale. Subject 5 and Subject 6 have the most significantly different ($p < 0.001$) features at the 5 min time scale.

3.1.2. Classification Accuracy of Different Classifiers with Different Time Scales. Figure 9 shows the classification accuracy of the mental workload using different classifiers with different time scales. Figure 9(a) shows the classification accuracy using SVM. It can be seen that the time scale with which the SVM achieved the highest average recognition accuracy was 3 min. In addition, the average classification accuracies of Subject 1 to Subject 6 with the 1, 3, and 5 min time scales were 95.30%, 97.54%, and 95.11%, respectively. Figure 9(b) shows the classification accuracy using KNN. It can be seen that the time scale with which the KNN obtained the highest average recognition accuracy was 3 min. In addition, the average classification accuracies of Subject 1 to Subject 6 with the 1, 3, and 5 min time scales were 96.09%, 98.77%, and 96.21%, respectively. Figure 9(c) shows the classification accuracy using GB; it achieved the highest average recognition accuracy with the 3 min time scale. In addition, the average classification accuracies of Subject 1 to Subject 6 with the 1, 3, and 5 min time scales were 93.17%, 95.90%, and 90.61%, respectively.

Figure 9(d) shows the classification accuracy using LDA; it did not achieve good classification performance with any of the three types of time scales. The average classification accuracies of Subject 1 to Subject 6 with the 1, 3, and 5 min time scales were 52.02%, 52.27%, and 52.28%, respectively. Figure 9(e) shows the classification accuracy using NB. It can be seen that NB achieved the highest average recognition accuracy with the time scale of 3 min. The average classification accuracies of Subject 1 to Subject 6 with the 1, 3, and 5 min time scales were 80.52%, 84.99%, and 80.07%, respectively. Finally, Figure 9(f) shows the classification accuracy using DT. The average classification accuracies of Subject 1 to Subject 6 with the 1, 3, and 5 min time scales were 80.52%, 84.99%, and 80.07%, respectively.

3.2. Cross-Subject Mental Workload Evaluation. The results of cross-subject mental workload evaluation at different time scales (1, 3, and 5 min) are presented in this subsection. The sample data of five of the six subjects were selected to train the machine learning model. At the same time, the sample data of the remaining subject were selected to test the machine learning model.

3.2.1. Statistically Significant Analysis of Features. Table 2 shows the statistical differences between the two groups at the time scales of 1, 3, and 5 min. From Table 2, we can see that there were 17 features in the most significantly different category ($p < 0.001$) and 2 features with significant differences ($p < 0.01$) between groups M-1 and R-1. There were eighteen features in the most significantly different category ($p < 0.001$) and two features of the significantly different category ($p < 0.01$) between groups M-3 and R-3. There were 17 features in the most significantly different category ($p < 0.001$) between groups M-5 and R-5.

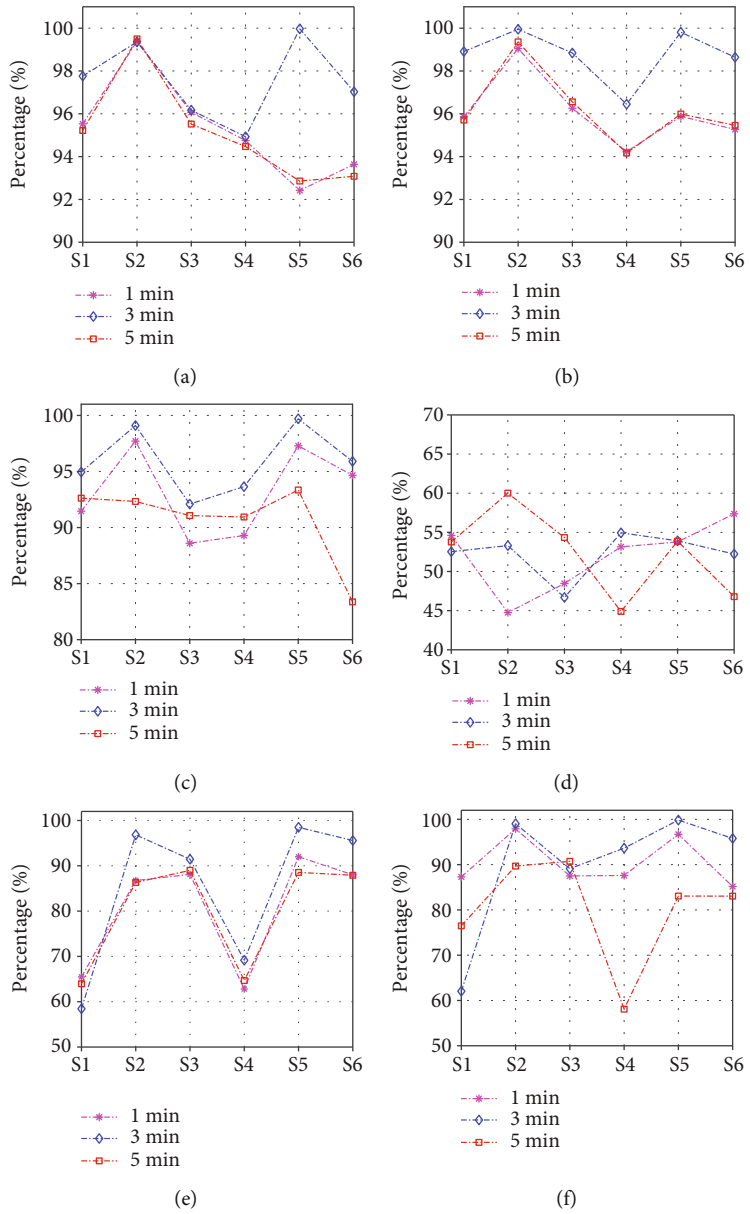

Figure 9: Classification accuracies of per-subject mental workload by different classifiers at different time scales: (a) support vector machine, (b) K-nearest neighbors, (c) gentle boost, (d) linear discriminant analysis, (e) naïve Bayes, and (f) decision tree.

3.2.2. Classification Accuracy of Different Classifiers with Different Time Scales. Figure 10 shows the classification accuracy of the mental workload using different classifiers at different time scales. Figure 10(a) shows the classification accuracy using SVM. It can be seen that when Subject 3 was used as the test subject, the classifier achieved the worst classification accuracy. The average classification accuracies of the classifier across all subjects with the 1, 3, and 5 min time scales were 77.59%, 75.06%, and 78.51%, respectively. Figure 10(b) shows the classification accuracy using KNN. Again, when Subject 3 was the test subject, the worst classification accuracy was achieved. The average classification accuracies of the classifier across all subjects with the 1, 3, and 5 min time scales were 69.24%, 70.40%, and 73.53%, respectively. Figure 10(c) shows the classification accuracy using GB. It can be seen that GB showed the worst classification accuracy with the time scale of 1 min and the best accuracy with the time scale of 5 min, both when Subject 2 was the test subject. The average classification accuracies of Subject 1 to Subject 6 with the 1, 3, and 5 min time scales were 63.53%, 71.55%, and 80.56%, respectively. Figure 10(d) shows the classification accuracy using LDA. It can be seen that the classifier showed the worst classification accuracy with the time scale of 3 min and the best accuracy with the time scale of 5 min, both when the data of Subject 3 were used as the test set. The average classification accuracies with the 1, 3, and 5 min time scales were 44.44%, 35.92%, and 53.92%, respectively. Figure 10(e) shows the classification accuracy using NB. It achieved the worst classification accuracy with Subject 3 as the test subject and the time scale of 5 min. It obtained the best accuracy with Subject 2 and the time scale of 5 min. The average classification accuracies

Table 2: Statistical analysis results of features under multiple time scales.

		M-1 and R-1	M-3 and R-3	M-5 and R-5
Time domain	HRVTi	0***	0***	0***
	Mean	0***	0***	0***
	SDNN	0***	0***	0***
	Median	0***	0***	0***
	pNN50	0***	0***	0***
	RMSSD	0***	0***	0***
Frequency domain	aHF	0***	0***	0***
	aLF	0***	0***	0***
	aTotal	0***	0***	0***
	aVLF	0***	0***	0***
	LF/HF	0***	0.054	0***
	nHF	0***	0***	0.001**
	nLF	0***	0***	0.002**
	pHF	0.80	0***	0***
	pLF	0.003**	0.60	0.194
	pVLF	0.006**	0***	0***
Nonlinear	SaEn	0***	0***	0***
	Alpha	0***	0***	0***
	Alpha1	0***	0***	0***
	Alpha2	0***	0***	0***

*, **, and *** represent $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively. M-1 : 1 min signals of the task-performing state; R-1 : 1 min signals of the rest state; M-3 : 3 min signals of the task-performing state; R-3 : 3 min signals of the rest state; M-5: 3 min signals of the task-performing state; R-5 : 5 min signals of the rest state.

with the 1, 3, and 5 min time scales were 64.53%, 66.48%, and 66.50%, respectively. Figure 10(f) shows the classification accuracy using DT. It can be seen that DT showed the worst classification accuracy with Subject 1 and the time scale of 5 min and the best accuracy with Subject 4 and the time scale of 5 min. The average classification accuracies with the 1, 3, and 5 min time scales were 65.03%, 67.91%, and 59.48%, respectively.

3.3. Discussion. Studies have shown that HRV can be used to measure and evaluate the mental workload of operators during human-robot interaction. Different time scales of HRV signals for mental workload measurement analysis have been widely studied. However, they were not based on a dataset of human-robot interaction. In addition, for the same dataset, the mental workload measurement analysis of human-robot interaction using HRV signals of different time scales was not reported, and there is no relevant public dataset. Hence, in this study, ECG signals were collected from six volunteers during task performance and rest. The fluctuation in the mental workload is closely related to the fluctuation state of the ANS, and HRV signals can react to the fluctuating state of the ANS. HRV signals of different lengths show levels of nervous activity information about the mental workload. This study presented a detailed comparative analysis.

First, the HRV signals at different time scales (1, 3, and 5 min) of the same individual were analyzed. Using a t-test, the statistical differences between the task-performing and rest states were analyzed. The results are shown in Figure 8. These are the p values of 1, 3, and 5 min time-scale HRV signals and the results with statistically significant features per subject at different time scales. It can be seen from Figure 8 that Subject 1 to Subject 4 show the most significantly different features at the 3 min time scale, whereas Subject 5 and Subject 6 have slightly less than the 5 min time scale. Moreover, there were a total of 75, 87, and 78 features with the most significant differences ($p < 0.001$) for the 1 min, 3 min, and 5 min time-scale HRV signals of the six subjects, respectively. It is shown that at the time scale of 3 min, there are more significantly different features than at the other time scales. The classification analysis of mental workload was performed using the features with statistical differences ($p < 0.05$) and six types of classifiers. The results are shown in Figure 9. It can be seen that the average accuracy across the six subjects with the 3 min time scale was the highest, i.e., 98.77% with the KNN classifier. The average accuracy across the six subjects at 1 min and 5 min were 96.09% (KNN) and 96.21% (KNN), respectively. This difference may be because the 1 min time-scale signal contains a limited amount of information. Although the 5 min time-scale signal contains a sufficient amount of information, the number of samples split from the collected signal is relatively small, which affects the training accuracy of the classification model. The signal length of 3 min contains sufficient time- and frequency-domain information, and more samples can be divided from the collected signals. Therefore, at a time scale of 3 min, the HRV signal analysis of the same individual obtained a high average classification accuracy. In addition, using 1, 3, and 5 min signals achieved high overall recognition accuracy and further verified that HRV signals can reflect the operator's mental workload changes during human-robot interaction.

HRV signals between different individuals were then analyzed. Using a t-test, the statistical differences between the task-performing and rest states were analyzed. The results are presented in Table 2. Table 2 shows that 17, 18, and 17 features were the most significantly different ($p < 0.001$) for 1 min, 3 min, and 5 min time-scale HRV signals of the six subjects, respectively. The classification analysis of mental workload was performed using the features with statistical differences ($p < 0.05$) and six types of classifiers. The sample data of five of the six individuals were used as the training set, and the sample data of one individual were left as the test set. The results are shown in Figure 10. It can be seen that the average accuracy of cross-subject identification is highest at 80.56% (GB) with the 5 min time scale, and the accuracies with 1 and 3 min time scales were 77.59% (SVM) and 75.06% (SVM), respectively. We found that the accuracy of cross-subject mental workload recognition was much lower than the per-subject mental workload recognition. This is because there are strong individual

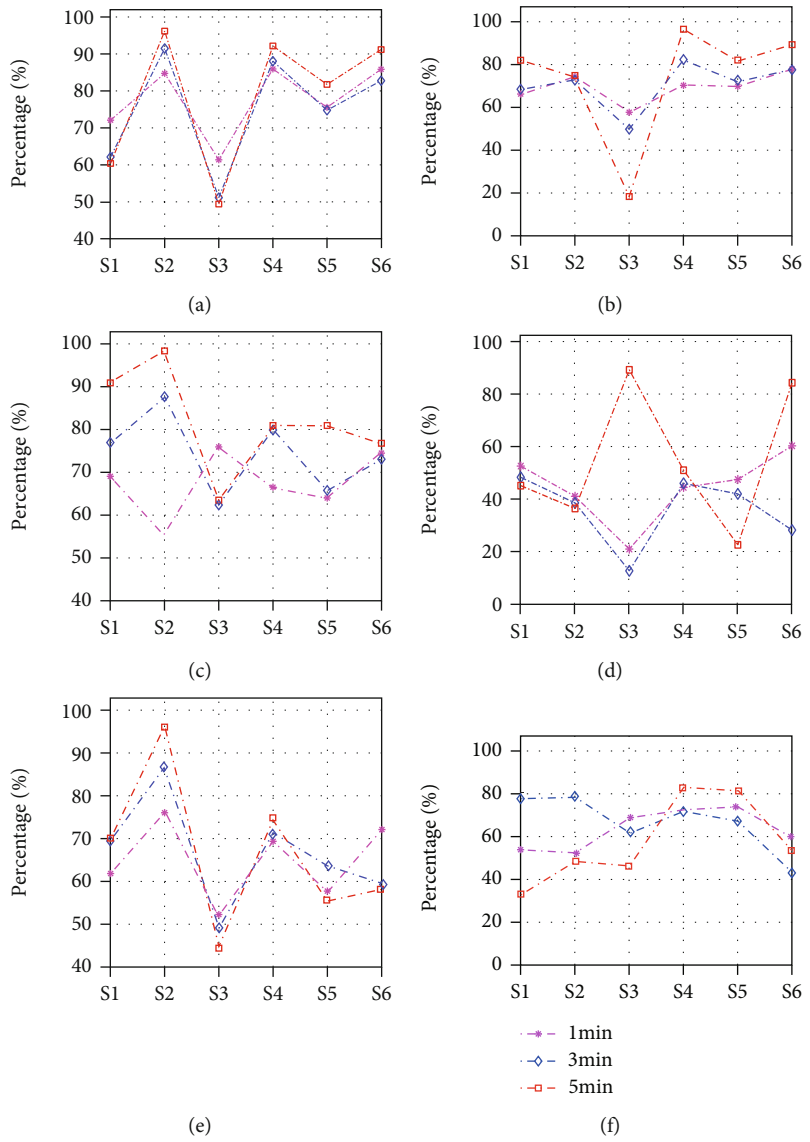

Figure 10: The classification accuracy of cross-subject mental workload by different classifiers at different time scales: (a) support vector machine, (b) K-nearest neighbors, (c) gentle boost, (d) linear discriminant analysis, (e) naïve Bayes, and (f) decision tree.

differences in HRV signals. Although HRV signals can reflect the fluctuating state of the ANS, there are differences in the psychological and physical qualities of different individuals. Therefore, to study cross-subject mental workload recognition, we need to further investigate the HRV signal to reflect the common characteristics of different individuals and to establish a universal mental workload recognition model.

4. Conclusion

In this paper, the differences in the recognition of the mental workload during human-robot interaction using multiple time-scale HRV signals were analyzed. First, ECG signals were obtained from six subjects while they were performing a task and while staying relaxed. Then, HRV signals were extracted based on the ECG signals. Furthermore, the HRV signals were divided into different groups using sliding windows of 1, 3, and 5 min. Then, several linear and nonlinear features of HRV signals were extracted for these different groups. Finally, six different machine learning algorithms were used to assess the mental workload performance. For the per-subject evaluation of mental workload with different time scales, the HRV signals of each individual were used for training, and then this individual's mental workload was assessed by the trained model. In the case of a 3 min signal length, the KNN method obtained an average accuracy of 98.77%. For the cross-subject mental workload evaluation, the HRV signals of five of six individuals were used to train the model. Then, the trained model identified the mental workload of the remaining individual. The highest average classification accuracy was obtained by the GB algorithm using the 5 min time scale, and its average accuracy was 80.56%. This study explores the problems of the operator's mental workload recognition during human-robot interaction using different time-scale HRV signals. However, the sample size in this study was limited; in the future, more data will be collected for analysis to provide

generalizable experimental results. In addition, online identification of human-robot interaction mental workload will be studied. Furthermore, different machine learning algorithms will be combined to choose the best recognition result of mental workload by voting.

Data Availability

Because the physiological signal of the human body involves personal privacy, so the experimental data will not be made public temporarily.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant number U20A20201), the Liaoning Province Doctoral Scientific Research Foundation (Grant number 2020-BS-025), the Liaoning Revitalization Talents Program (Grant number XLYC1807018), and the National Key Research and Development Program of China (Grant number 2016YFE0206200).

References

[1] A. Costes, F. Danieau, F. Argelaguet, P. Guillotel, and A. Lecuyer, "Towards haptic images: a survey on touchscreen-based surface haptics," *IEEE Transactions on Haptics*, vol. 13, no. 3, pp. 530–541, 2020.

[2] S. Shao, T. Wang, Y. Su, C. Yao, C. Song, and Z. Ju, "Multi-IMF sample entropy features with machine learning for surface texture recognition based on robot tactile perception," *International Journal of Humanoid Robotics*, vol. 18, no. 2, p. 2150005, 2021.

[3] W. Zheng, H. Liu, and F. Sun, "Lifelong visual-tactile cross-modal learning for robotic material perception," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 3, pp. 1192–1203, 2021.

[4] P. Falco, S. Lu, C. Natale, S. Pirozzi, and D. Lee, "A transfer learning approach to cross-modal object recognition: from visual observation to robotic haptic exploration," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 987–998, 2019.

[5] Jongdae Jung, Seung-Mok Lee, and Hyun Myung, "Indoor mobile robot localization and mapping based on ambient magnetic fields and aiding radio sources," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 7, pp. 1922–1934, 2015.

[6] W. Yuan, Z. Li, and C.-Y. Su, "Multisensor-based navigation and control of a mobile service robot," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 4, pp. 2624–2634, 2021.

[7] E. Debie, R. Fernandez Rojas, J. Fidock et al., "Multimodal fusion for objective assessment of cognitive workload: a review," *IEEE Transactions on Cybernetics*, vol. 51, no. 3, pp. 1542–1555, 2021.

[8] Z. Pei, H. Wang, A. Bezerianos, and J. Li, "EEG-based multiclass workload identification using feature fusion and selection," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, article 4001108, 2021.

[9] S. Delliaux, A. Delaforge, J.-C. Deharo, and G. Chaumet, "Mental workload alters heart rate variability, lowering nonlinear dynamics," *Frontiers in physiology*, vol. 10, article 565, 2019.

[10] K. Tsunoda, A. Chiba, K. Yoshida, T. Watanabe, and O. Mizuno, "Predicting changes in cognitive performance using heart rate variability," *IEICE Transactions on Information and Systems*, vol. E100.D, no. 10, pp. 2411–2419, 2017.

[11] S. Shao, T. Wang, C. Song, X. Chen, E. Cui, and H. Zhao, "Obstructive sleep apnea recognition based on multi-bands spectral entropy analysis of short-time heart rate variability," *Entropy*, vol. 21, no. 8, p. 812, 2019.

[12] S.-L. Shao, T. Wang, C.-H. Song, E. N. Cui, H. Zhao, and C. Yao, "A novel method of heart rate variability measurement," *Acta Physica Sinica*, vol. 68, no. 17, article 178701, 2019.

[13] A. Tiwari, I. Albuquerque, M. Parent et al., "Multi-scale heart beat entropy measures for mental workload assessment of ambulant users," *Entropy*, vol. 21, no. 8, p. 783, 2019.

[14] A. Tiwari, S. Narayanan, and T. H. Falk, "Stress and anxiety measurement "In-the-Wild" using quality-aware multi-scale HRV features," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 7056–7059, Montreal, Canada, 2019.

[15] P. Gilfriche, L. M. Arsac, Y. Daviaux et al., "Highly sensitive index of cardiac autonomic control based on time-varying respiration derived from ECG," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 315, no. 3, pp. R469–R478, 2018.

[16] M. Malik, J. T. Bigger, A. J. Camm et al., "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *European Heart Journal*, vol. 17, no. 3, pp. 354–381, 1996.

[17] V. Villani, M. Righi, L. Sabattini, and C. Secchi, "Wearable devices for the assessment of cognitive effort for human–robot interaction," *IEEE Sensors Journal*, vol. 20, no. 21, pp. 13047–13056, 2020.

[18] S. Shao, T. Wang, Y. Wang, Y. Su, C. Song, and C. Yao, "Research of HRV as a measure of mental workload in human and dual-arm robot interaction," *Electronics*, vol. 9, no. 12, article 2174, 2020.

[19] V. Villani, B. Capelli, C. Secchi, C. Fantuzzi, and L. Sabattini, "Humans interacting with multi-robot systems: a natural affect-based approach," *Autonomous Robots*, vol. 44, pp. 601–616, 2020.

[20] R. Castaldo, L. Montesinos, P. Melillo, C. James, and L. Pecchia, "Ultra-short term HRV features as surrogates of short term HRV: a case study on mental stress detection in real life," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, article 12, 2019.

[21] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. -BME-32, no. 3, pp. 230–236, 1985.

[22] L. Chen, X. Zhang, and C. Song, "An automatic screening approach for obstructive sleep apnea diagnosis based on single-lead electrocardiogram," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 1, pp. 106–115, 2015.

[23] F. Zhu, J. Yang, S. Xu, C. Gao, N. Ye, and T. Yin, "Incorporating neighbors' distribution knowledge into support vector

machines," *Soft Computing*, vol. 21, no. 21, pp. 6407–6420, 2017.

[24] F. Zhu, Y. Ning, X. C. Chen, Y. Zhao, and Y. Gang, "On removing potential redundant constraints for SVOR learning," *Applied Soft Computing*, vol. 102, no. 4, article 106941, 2021.

[25] Y. Li, W. Pan, K. Li, Q. Jiang, and G. Liu, "Sliding trend fuzzy approximate entropy as a novel descriptor of heart rate variability in obstructive sleep apnea," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 1, pp. 175–183, 2019.

[26] K. Machetanz, L. Berelidze, R. Guggenberger, and A. Gharabaghi, "Brain-heart interaction during transcutaneous auricular vagus nerve stimulation," *Frontiers in Neuroscience*, vol. 15, article 632697, 2021.

[27] G. D. Clifford and L. Tarassenko, "Quantifying errors in spectral estimates of HRV due to beat replacement and resampling," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 4, pp. 630–638, 2005.

[28] W. Zheng, S. Chen, Z. Fu, F. Zhu, H. Yan, and J. Yang, "Feature selection boosted by unselected features," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2021.