Sensors and Materials, Vol. 38, No. 3 (2026) 1207–1227
MYU Tokyo

S & M 4372

Mental Workload Estimation During Floor Cleaning Based on Wearable Inertial Sensors

Moemi Shidahara[1] and Kaori Fujinami[1,2*]

[1]Graduate School of Bio-Applications and Systems Engineering, Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei, Tokyo 184-8588, Japan
[2]Division of Advanced Information Technology and Computer Science, Institute of Engineering, Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei, Tokyo 184-8588, Japan

(Received September 9, 2025; accepted December 26, 2025)

Keywords: mental workload, NASA-TLX, behavioral data, inertial sensors, symbol sequence

Mental workload (MWL) is the cognitive effort required to manage information in working memory. Excessive MWL increases the error risk and, when prolonged, can impair appetite, sleep, and overall health. Therefore, an objective and real-time MWL estimation is crucial. In this study, we introduce an MWL estimation method during floor-cleaning tasks using inertial sensor data collected from the body and cleaning tools. We introduce conventional statistical features from inertial sensor signals and two types of feature derived from symbol sequences via vector quantization. We construct regression models to estimate MWL and compare their errors using various combinations of these three feature types. The models consistently achieve lower errors than a naive baseline, which always predicts the training data median. We also compare results from different perspectives, such as sensor placement in each scenario and the computation time required for feature extraction. The findings suggest that the proposed approach has practical potential for the daily monitoring and visualization of MWL.

1. Introduction

Accidents caused by human error are frequently reported in aviation,[(1)] driving,[(2)] and other fields. Under multitasking conditions, the mental burden imposed on working memory is referred to as cognitive load.[(3)] Mental workload (MWL) is a quantitative representation of this cognitive load and is closely related to human performance and behavior.[(4)] Given this strong relationship, MWL is considered a critical measure to ensure operational safety.[(5)] In addition to its relationship with performance and safety, MWL has also been linked to health outcomes. Chaput and Tremblay reported that voluntary food intake significantly increases in individuals with a high MWL.[(6)] Cropley *et al*. cited evidence indicating that chronic workload is associated with a heightened risk of fatigue and sleep disorders,[(7)] suggesting that MWL monitoring can be valuable in healthcare contexts. Haapalainen *et al*. emphasized the importance of estimating MWL in real time and applying it in daily life settings.[(8)]

*Corresponding author: e-mail: fujinami@cc.tuat.ac.jp
https://doi.org/10.18494/SAM5933

ISSN 0914-4935

In this study, we aim to quantitatively estimate MWL from behavioral measurements obtained from daily activities. We also aim to develop an MWL estimation method that enables long-term sensing in a real-world setting using wearable sensors. Uemura *et al*. employed accelerometers for MWL estimation;[9] however, requiring users to wear a specific device around the waist can be cumbersome, limiting its applicability for extended use. Consequently, leveraging devices that are already worn or routinely used in daily life may help reduce the burden of wearing additional equipment. Moreover, for behavioral sensing to be sustainable over long durations, the method should be applicable not only in constrained environments such as offices but also across a diverse range of everyday activities, including household chores. In daily routines, people interact with various tools such as vacuum cleaners and serving trays. Detecting the motion of such tools may broaden the applicability of the MWL estimation method to a wider range of daily tasks. In this study, we first focus on a particular daily activity to verify the feasibility of the MWL estimation method, with the long-term aim of generalizing the method to other tasks. The selected activity is required to meet the following criteria:

(a) the use of a tool to which a sensor can be attached,
(b) the involvement of full-body physical movement, and
(c) a continuous duration from start to finish.

We developed an MWL estimation method for an activity that satisfies these conditions, specifically floor cleaning using a flooring wiper (floor cleaning). When performing this task, behavioral data, including acceleration and angular velocity, are collected, and the subjective MWL is assessed using a questionnaire. We then constructed a regression model using the features extracted from the sensor data as explanatory variables and the subjective MWL score as the objective variable. This approach enables the quantitative estimation of MWL based on behavioral sensing.

We investigated key design parameters that affect the performance of the regression model, such as window size and regression model, to improve the accuracy (estimation error) of the MWL estimation system. Because the system is intended for long-term use in daily life, we hypothesize that capturing patterns from long-term behavioral transitions can further improve the estimation accuracy. In our previous work,[10] we segmented collected behavioral data into fixed-length windows and computed statistical features such as the mean and standard deviation (*SD*) within each window. These are referred to as the basic features. In this study, we focus on behavioral transitions by converting the multidimensional information derived from behavioral data into symbol sequences using vector quantization. We also introduce two types of feature that characterize the properties of these symbol sequences.

The remainder of this paper is organized as follows. In Sect. 2, we review the related work. In Sect. 3, we define the system requirements and configuration, and describe the features used for MWL estimation. Data collection for constructing and testing the MWL estimation model is presented in Sect. 4. Experiments to understand the effects of various parameters in terms of accuracy and computational cost are presented in Sect. 5, followed by a discussion in Sect. 6. Finally, in Sect. 7, we conclude the paper and outline directions for future research. The Tokyo University of Agriculture and Technology Ethics Committee approved this study (approval number: 240228-0571).

2. Related Work

2.1 Estimation of human internal states

In the work on estimating internal states such as MWL, subjective evaluations using questionnaires have traditionally been employed;[2] however, administering questionnaires during or immediately after a task can ironically increase the MWL itself because the cognitive resources required to respond to pop-up prompts and retain response content may interfere with the primary task.[11,12] Consequently, we aim to passively and unobtrusively estimate the internal states in the background during task execution. This approach relies on objectively measurable information, such as biometric signals [e.g., pulse waves and electroencephalograms (EEGs)] and behavioral data derived from body movements, as indicators of internal states.

Shimizu *et al*. assessed the emotional states of automobile drivers using biometric signals such as EEGs and pulse rates.[13] They focused on classifying emotions as positive, negative, or neutral. Similarly, Tanaka *et al*. investigated the recognition of three internal states—concentration, inhibition, and fatigue—during a silent reading task using eye tracking technology.[14] They reported an F1-score of 0.637 and identified effective feature sets for distinguishing these states. These findings suggest that internal states can be inferred by clarifying their associations with the numerical values collected using sensors. Estimating the internal states using such quantitative data enables objective evaluation and holds promise for effective MWL estimation.

2.2 MWL estimation

2.2.1 NASA-Task Load Index (NASA-TLX)

The NASA-TLX assesses MWL across six distinct dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration level.[15] The importance (weight) of each dimension is determined on the basis of user judgment through pairwise comparison, and a workload score ranging from 0 to 100 is calculated as the weighted average of the ratings multiplied by their corresponding weights.[2,12,16]

2.2.2 Use of biometric sensors

Various biometric signals, such as EEGs and electrocardiograms (ECGs), have been used as objective indicators to estimate MWL. Causse *et al*. evaluated MWL and psychological stress during an actual flight using the heart rate data derived from ECGs combined with subjective MWL ratings.[1] They observed a strong positive correlation between heart rate and MWL, and a similar trend was also observed for psychological stress. These findings demonstrate the effectiveness of biometric signal measurements for estimating workload and stress in real flight environments. Conway *et al*. conducted an experimental study to examine the effect of stress on cognitive load indicators derived from the galvanic skin response (GSR).[17] The results

demonstrated that the mean GSR became less reliable under stress, whereas the peak frequency and duration remained stable and were effective as indicators of cognitive load.

Although the MWL estimation method based on biometric signals has shown effectiveness, it is often constrained to binary classification tasks, such as detecting the presence or absence of a workload. This limits its applicability to the quantitative estimates required in this study.[12] Moreover, biometrics-based approaches typically require dedicated equipment and controlled environments, which restrict their practicality in everyday contexts.[8] This challenge becomes even more pronounced in real-world scenarios, where such systems are highly susceptible to external disturbances. Although the use of environment-independent wearable sensors such as smartwatches has become increasingly common, physical activity often introduces motion-induced artifacts that are difficult to disentangle from the physiological signals themselves, making it infeasible to directly use the original signals as reliable indicators for MWL estimation.[12]

2.2.3 Use of nonbiometric information sources

A methodology has been proposed for the quantitative estimation of MWL using machine learning (ML) that incorporates both objective indicators, such as voice characteristics and hand and body movements, and subjective evaluations, such as NASA-TLX, as the ground truth. Chua *et al.* constructed a classification model to identify users' arousal and cognitive load in a virtual reality environment by analyzing their hand and head movements. They observed a notable decrease in the speed of hand movement, an increase in hand tension, and a reduction in head motion.[18] Similarly, Chen *et al*. estimated the cognitive load during a memory task by analyzing behavioral features, such as eye movements, mouse and keyboard interactions, voice characteristics, and behavioral patterns, along with biometric signals, subjective ratings, and task performance scores.[19] These studies suggest that behavioral data can serve as effective objective indicators for MWL estimation.

Furthermore, methods employing wearable sensors have been proposed for acquiring behavioral data. Uemura *et al*. focused on walking as an automatic task typically performed unconsciously in daily life and hypothesized that a higher cognitive load leads to greater distraction and reduced focus on physical tasks that can be performed by motor reflex, such as walking.[9] Their findings revealed that the amplitude of angular velocity tended to decrease under elevated load conditions, suggesting the feasibility of estimating the cognitive load using only wearable sensors.

These results support the efficacy of the MWL estimation method using wearable and stationary sensors, in conjunction with behavioral data. Unlike physiological signal-based methods, this approach imposes fewer constraints on physical movement, thereby enhancing its applicability to real-world daily scenarios.

3. MWL Estimation Method

3.1 Equipment and materials

As illustrated in Fig. 1, the system we developed uses two inertial measurement units (IMUs) (ATR Promotions Inc., TSND151[20]) and an earphone-type device (eSense[21]) equipped with an IMU to acquire motion data from the head, (dominant) wrist, and wiper. The mounting positions and coordinate axes of the TSND151 sensors are illustrated in Fig. 2. On the wiper, a thin plate was fixed to the side surface approximately 20.5 cm from the tip using screws, and the sensor was attached to the plate using Velcro tape. The sensor was placed on the wrist in a small pouch attached to a wristband fastened with a Velcro-brand product. The coordinate axes of both sensors were aligned as shown in Fig. 2. No additional calibration procedures, such as gravity alignment, static pose calibration, drift correction, and axis reorientation, were performed. Each sensor's intrinsic coordinate system, defined by its mounting orientation, was used without transformation to a global reference frame.

All three sensors were connected to a single PC via Bluetooth. The two TSND151 units were synchronized to the PC's internal clock at the beginning of each task to collect the data sampled at 50 Hz, whereas the eSense device sampled at 5 Hz and was timestamped on the PC upon reception. Although the sampling rates differed, we evaluated the MWL estimation performance of each sensor independently. Therefore, no interpolation or resampling was performed. To maintain temporal consistency within each analysis window, the first eSense sample closest to the first TSND151 sample was designated as the starting point of the window. This limited the maximum window offset to the eSense sampling interval (approximately 200 ms, about 4% of the 4.8 s window), which indicates that the system was nearly synchronized at the window level across sensors.

The acquired data comprise seven-dimensional signals: three-axis gravitational acceleration, three-axis angular velocity, and composite acceleration. The behavioral data were segmented using a fixed-length sliding window approach, and three types of feature were extracted from each window. The extracted features are used as explanatory variables in the regression model.

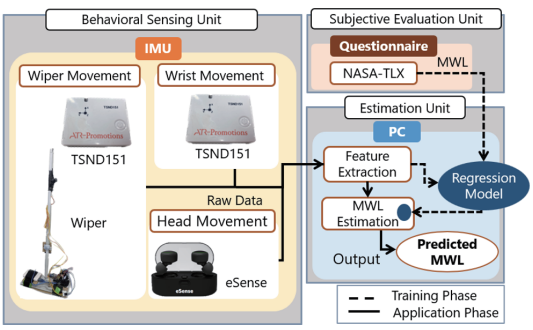

Fig. 1. (Color online) System overview.

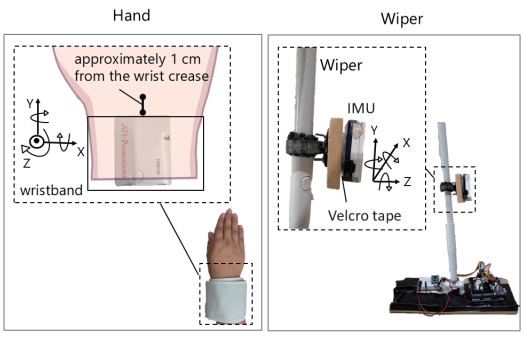

Fig. 2. (Color online) Mounting positions and axes of TSND151 sensors on the wrist and wiper.

The objective variable for the regression is the weighted workload score obtained from the NASA-TLX, which represents the level of MWL.

3.2 Feature engineering

3.2.1 Exploration of features

Figure 3 presents examples of the behavioral data collected during the experiment, along with the corresponding MWL labels. In this figure, gyrx, gyry, and gyrz denote the angular velocity components around the *x*-, *y*-, and *z*-axes of the head, respectively. The waveforms of these components are compared under the two conditions for a specific subject: one with a relatively low MWL (low-load, MWL: 6.9) and one with a relatively high MWL (high-load, MWL: 30.9). The waveform amplitude is larger under the low-load condition and smaller under the high-load condition. On the basis of this observation, we hypothesized that the amplitudes of behavioral signals can serve as indicators of MWL.

Time-series signals were converted into symbol sequences by vector quantization to further analyze the temporal structure of behavioral data. For instance, the waveform in Fig. 3 was encoded as "ACDDE" under the low-load condition and "BCBBB" under the high-load condition. From these sequences, a weighted directed network was constructed (Fig. 4), where each symbol represents a node, transitions between symbols are represented as edges, and transition probabilities are used as edge weights. As illustrated in Fig. 4, the low-load condition exhibited more diverse transitions, whereas the high-load condition exhibited more repetitive transitions. On the basis of this observation, we hypothesized that the complexity of the transition network is associated with MWL.

In addition, the bag-of-symbols (BoS) method, inspired by the bag-of-words model used in natural language processing,[22] was applied to symbol sequences. Symbol occurrence probabilities were calculated and visualized as normalized histograms in Fig. 5, where the distribution of a single symbol, that is, a unigram, differed under MWL conditions. Moreover, we can consider two and three consecutive symbols, known as bigrams and trigrams, respectively.

In total, three types of feature were utilized for MWL estimation: basic features directly obtained as descriptive statistics from the sensor data, network (NW) features obtained from symbol transition networks, and BoS features representing symbol occurrence patterns. Among these, the NW and BoS features are collectively referred to as symbol-sequence features.

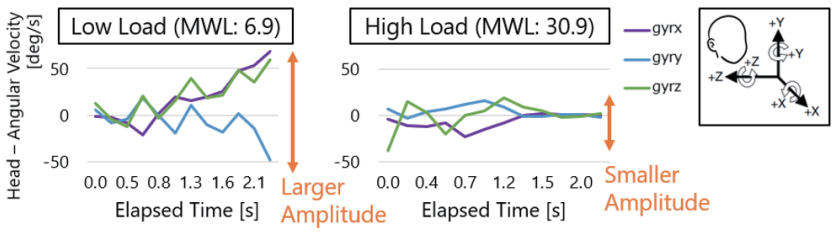

Fig. 3. (Color online) Example of behavioral data and corresponding MWL labels.

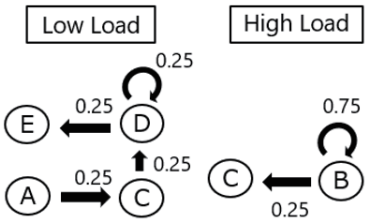

Fig. 4. Example of symbol transition networks as weighted directed graphs.

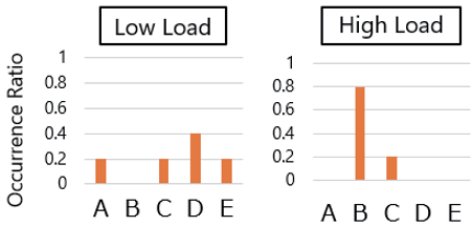

Fig. 5. Example of symbol occurrence histograms.

3.2.2 Definition of features

The basic features comprised ten time-domain features, that is, mean, *SD*, maximum, minimum, range, variance, median, skewness, kurtosis, and mean absolute deviation, and four frequency-domain features, that is, peak frequency, power spectral density at the peak frequency, energy, and entropy. As each IMU provides seven-dimensional data, 98 basic features were computed.

The process of generating symbol-sequence features via vector quantization is illustrated in Fig. 6. In vector quantization, a set of multidimensional data (i.e., a vector) is assigned to a specific symbol. The aforementioned time-domain features were used as a vector to be quantized, which is calculated for each ω-sized sliding window with 50% overlap, that is, a shift of $\omega/2$, and referred to as symbolization features. Before assigning symbols to the vectors, the vector space is partitioned into N areas by clustering, implying that a cluster centroid serves as a representation of a symbol. We used 5% of the randomly sampled vectors from all collected data for clustering. In the operation phase, a sequence of vectors is encoded into a sequence of symbols by assigning each vector to the nearest centroid. ω and N are hyperparameters and optimized via grid search to maximize MWL estimation performance.

The extraction procedures for the NW and BoS features from the symbol sequences are shown in Fig. 7. Each symbol sequence was divided into S segments to obtain the NW and BoS features. The NW features were defined by treating symbols as nodes in a network, where the measures used in network sciences were applied: five node-level measures (in-degree, out-degree, closeness centrality, betweenness centrality,[23] and PageRank[24]); for in-degree, out-degree, and PageRank, *SD* was calculated; and for closeness centrality and betweenness centrality, both the mean and *SD* were computed. Additionally, two global measures (network density and network transitivity[23]) were defined. To this end, measures with a total of $5N + 9$ $[= 3(N + 1) + 2 (N + 2) + 2]$ dimensions were defined as NW features.

The dimensionality of the BoS features was N, N^2, and N^3 for unigrams, bigrams, and trigrams, respectively. In the subsequent evaluation, the effectiveness of these three feature sets was comparatively evaluated on the basis of their regression performance in MWL estimation.

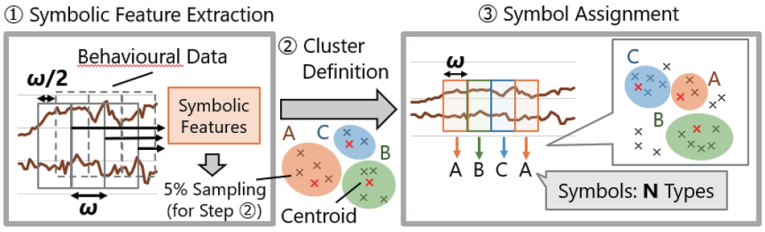

Fig. 6. (Color online) Feature extraction process through symbolization of behavioral data.

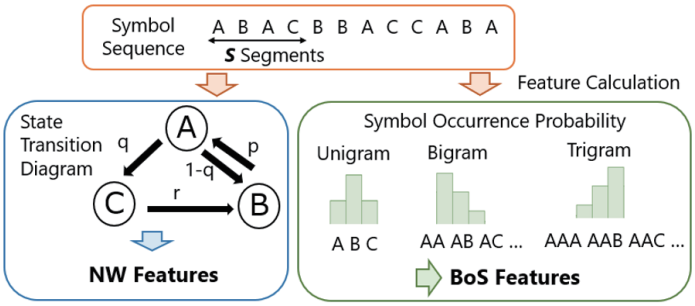

Fig. 7. (Color online) Generating NW and BoS features from symbol sequences.

4. Data Collection

4.1 Method

We created a dataset to construct a regression model to estimate MWL during floor cleaning tasks. A total of 12 participants (six males and six females) in their 20s participated in the study. The participants were instructed to clean a predefined area by following a straight or curved path on the floor (Fig. 8). Data were collected for three different surface materials where the body and wiper movements varied: (a) simili paper, (b) simili paper on corrugated paper, and (c) glossy paper. To collect data on diverse cognitive loads, participants completed three types of secondary task with various cognitive loads, as described in Sect. 4.2. One cleaning task with a specific secondary task comprised five sequences in which each sequence arranged the surface materials in the following order: (a)→(b)→(c)→(b)→(a)→(a). The required time was approximately 5 min. We used six different secondary task orders across participants to counterbalance the order effects. Before the first task, the participants had time to become familiar with the sensing devices and review the designated cleaning routes and procedures. A 10 min break was given between tasks. At the end of each task sequence, the participants answered the NASA-TLX questionnaire, rating the six subscales from 0 to 100. These subjective ratings were used as the ground truth for MWL. All data were collected with informed consent from the participants and are not publicly available in compliance with the consent agreements.

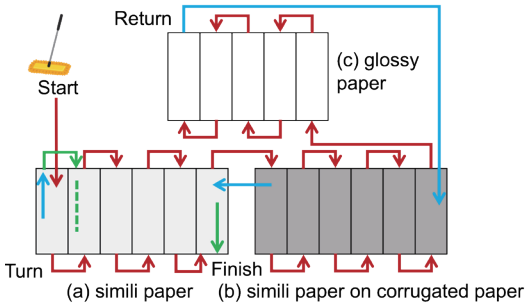

Fig. 8. (Color online) Floor layout and cleaning path in data collection.

4.2 Establishing the level of work difficulty

Memory search was adopted as a secondary task in data collection. As summarized in Table 1, three levels of difficulty were defined by combining the primary task (i.e., floor cleaning) with a memory task. The subtask involved six auditory stimuli. In the easy memory task, the participants were instructed to memorize one specific sound, whereas in the hard task, they memorized two sounds. Upon hearing a memorized sound, they tilted the joystick (ELEGOO Joystick module[25]) attached to the wiper to the right; for nontarget sounds, they tilted it to the left.

5. Evaluation

5.1 Verification method

From the collected behavioral data, three types of feature were calculated as explanatory variables to construct a regression model using ML. The development and evaluation environments are listed in Table 2. In the modeling process, the time window W for feature calculation was set to 240 samples (4.8 s), based on prior validation,[10] indicating that this window length yielded the highest MWL estimation accuracy, and a sliding window with 50% overlap (120 samples or 2.4 s) was adopted. A light gradient boosting machine [LightGBM (LGBM)] was employed on the basis of prior evaluation,[10] which demonstrated the highest performance. The hyperparameters of the LGBM model used in this study are summarized in Table 3. All unspecified hyperparameters in the LGBM library were set to their default values.

The regression model was evaluated using the mean absolute error (MAE) defined in Eq. (1), where $\hat{y}_i$, y_i, and m denote the predicted value, the observed value (ground truth), and the number of test samples, respectively. A lower MAE indicates a higher estimation performance. Tenfold cross-validation was performed per participant, and an average MAE was reported, which represented the average performance when the estimation model was customized for each participant. In this evaluation, random sampling was not used to test the model; instead, the test data were selected in their original temporal order. This setting reflects a realistic usage scenario in which the data newly collected during system operation are temporally separated from the

Table 1
Task settings for different levels of difficulty.

Level of difficulty	Task
Low	Floor cleaning only
Middle	Floor cleaning and easy memory task
High	Floor cleaning and hard memory task

Table 2
Development and execution environments.

CPU	Intel Xeon W-2223
Memory	32 GB
OS	Windows 11 Pro for Workstation
Programming language	Python 3.11.7
Libraries	scikit-learn 1.2.2, LightGBM 4.3.0

Table 3
LGBM hyperparameters.

Item	Setting/status
Metric	L2 loss
Learning rate	0.1
Max depth	Not restricted (−1)
Num leaves	31
Subsampling	Not applied
Early stopping	Not applied

training data and may exhibit different trends. Therefore, this approach was adopted to more accurately approximate the prediction performance expected in actual usage. In addition, feature standardization was performed within each training fold by z-score normalization based only on the training data to prevent data leakage.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|\hat{y}_i - y_i\right| \tag{1}$$

All features were calculated using the same window to ensure a fair comparison across feature types. The number of samples per symbol, ω, and the number of segments, S, used for BoS and NW feature computation were determined such that they satisfy Eq. (2), where $floor(\cdot)$ denotes the floor function, that is, rounding down. The parameters were optimized to ensure a valid temporal segmentation.

$$W = \omega \times floor((S + 1)/2) \tag{2}$$

Unless otherwise noted, the evaluations were based on the wrist-worn sensor data, which demonstrated the highest estimation accuracy among the three sensor positions (lowest *MAE*).

5.2 Parameter selection for symbol sequence features

The parameters, namely, ω, S, and N, may affect both the accuracy of MWL estimation and the associated computational cost in the symbol sequence features. The performance may further depend on the feature extraction method, such as that involving unigrams, bigrams, or trigrams, in the BoS features. Thus, we compared various combinations of these parameters to understand the performance variations in *MAE* and computation time.

5.2.1 Comparison by estimation error

Five values of ω (4, 8, 12, 20, and 24 samples) were specified, and five values of S (120, 60, 40, 24, and 20 samples) were obtained accordingly by applying Eq. (2) with $W = 240$. All combinations of these values were evaluated systematically. Considering that the sampling frequency of the head-mounted sensor differs from those of the wrist- and wiper-mounted sensors, three specific values of ω (4, 8, and 12) and three values of S (12, 6, and 4) were selected. To further investigate the effect of the number of symbols, N was varied across 5, 10, 15, and 20 for each ω–S combination, and the estimation errors were calculated accordingly. For the BoS features, the evaluation was performed according to the number of consecutive symbols, that is, unigram, bigram, and trigram.

Table 4 shows the results of NW features. For BoS features, the results corresponding to the unigrams, bigrams, and trigrams are summarized in Tables 5–7, respectively. In all the tables, the highest-performing configuration (i.e., the configuration with the lowest *MAE*) is marked with an underline. Values in parentheses represent *SD* across participants. The results indicate that the combination of $\omega = 4$ and $N = 20$ consistently yielded the lowest estimation error, regardless of the feature type or BoS feature calculation method. This trend was also observed for the other sensor positions. Furthermore, as shown in Tables 5–7, both bigrams and trigrams outperformed unigrams in terms of estimation error. The bigrams and trigrams achieved significantly lower *MAE*s ($p < 0.05$) than the unigrams as a result of the Wilcoxon signed-rank test, followed by Ryan's multiple comparison test.

5.2.2 Comparison by computational cost

A substantial increase in the computational cost of feature calculation can negatively affect the usability of the MWL estimation system and lead to increased power consumption. Thus, the feature calculation time was compared across different parameter values and feature calculation methods. The feature calculation time was defined as the duration required to compute each feature vector used as an input to the regression model. The time required to train the regression model was excluded. The measured portion of the process in the pipeline differs between the basic and symbolic sequence features, as indicated by the dotted arrows in Fig. 9.

The comparison focused on the effect of two parameters, ω and N, as well as the number of consecutive symbols in the BoS feature. First, ω was varied while N was fixed at 20, and the feature computation time was measured. The results are summarized in Table 8. As indicated,

Table 4
Effect of the combination of N and ω on *MAE* in NW features.

	$N = 5$	$N = 10$	$N = 15$	$N = 20$
$\omega = 4$	9.8 (4.1)	9.5 (3.8)	9.3 (3.7)	9.1 (3.7)
$\omega = 8$	9.8 (4.1)	9.6 (3.9)	9.5 (3.8)	9.4 (3.7)
$\omega = 12$	9.7 (4.1)	9.6 (3.8)	9.5 (3.8)	9.5 (3.8)
$\omega = 20$	9.7 (4.2)	9.7 (3.9)	9.6 (3.8)	9.6 (3.9)
$\omega = 24$	9.7 (4.1)	9.7 (3.9)	9.7 (3.9)	9.5 (3.7)

Table 5
Effect of the combination of N and ω on *MAE* in BoS features (unigram).

	$N = 5$	$N = 10$	$N = 15$	$N = 20$
$\omega = 4$	9.8 (4.1)	9.5 (3.8)	9.3 (3.7)	9.1 (3.7)
$\omega = 8$	9.8 (4.1)	9.6 (3.9)	9.5 (3.8)	9.4 (3.7)
$\omega = 12$	9.7 (4.1)	9.6 (3.8)	9.5 (3.8)	9.5 (3.8)
$\omega = 20$	9.7 (4.2)	9.7 (3.9)	9.6 (3.8)	9.6 (3.9)
$\omega = 24$	9.7 (4.1)	9.7 (3.9)	9.7 (3.9)	9.5 (3.7)

Table 6
Effect of the combination of N and ω on *MAE* in BoS features (bigram).

	$N = 5$	$N = 10$	$N = 15$	$N = 20$
$\omega = 4$	9.7 (3.8)	9.3 (3.7)	9.1 (3.5)	8.9 (3.5)
$\omega = 8$	9.7 (3.9)	9.4 (3.7)	9.3 (3.7)	9.2 (3.6)
$\omega = 12$	9.7 (3.9)	9.4 (3.7)	9.3 (3.6)	9.2 (3.7)
$\omega = 20$	9.7 (4.0)	9.6 (3.8)	9.5 (3.8)	9.4 (3.8)
$\omega = 24$	9.7 (4.0)	9.6 (3.8)	9.6 (3.9)	9.3 (3.6)

Table 7
Effect of the combination of N and ω on *MAE* in BoS features (trigram).

	$N = 5$	$N = 10$	$N = 15$	$N = 20$
$\omega = 4$	9.6 (3.7)	9.1 (3.6)	9.1 (3.5)	8.9 (3.5)
$\omega = 8$	9.7 (3.9)	9.4 (3.6)	9.3 (3.6)	9.4 (3.7)
$\omega = 12$	9.7 (3.9)	9.4 (3.6)	9.3 (3.6)	9.3 (3.7)
$\omega = 20$	9.7 (4.0)	9.7 (3.8)	9.6 (3.8)	9.6 (3.7)
$\omega = 24$	9.7 (4.0)	9.7 (3.8)	9.6 (3.9)	9.5 (3.7)

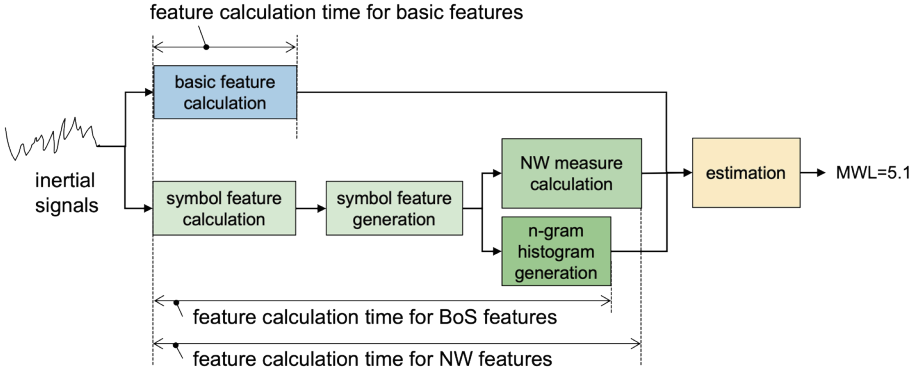

Fig. 9. (Color online) Overview of the estimation model process and segments related to feature calculation.

Table 8
Relationship between ω and feature computation time (s) ($N = 20$).

	$\omega = 4$	$\omega = 8$	$\omega = 12$	$\omega = 20$	$\omega = 24$
NW features	0.04	0.04	0.05	0.07	0.15
BoS features	0.01	0.01	0.02	0.05	0.09

the computation time was highly sensitive to ω, with the smallest window size ($\omega = 4$) requiring approximately 3.8 times more time than the largest window size ($\omega = 24$). Next, N was varied while maintaining $\omega = 4$ fixed, and the results are summarized in Table 9. Finally, feature computation times were compared across different BoS feature types. As summarized in Table 10, unigrams and bigrams required similar computation times, whereas trigrams significantly increased the processing time.

5.3 Comparison between individual feature types

The effectiveness values of the three feature types were compared. The parameters combined for the symbol sequence features were $\omega = 4$, $S = 120$, and $N = 20$, which yielded the lowest estimation error in Sect. 5.2. A bigram was selected for the BoS features. To assess the regression performance of each feature type, the *MAE* of the baseline regressor was compared with that of LGBM. The baseline regressor always returns the median of the training data and is implemented in scikit-learn as DummyRegressor (with the parameter strategy = "median"), which is abbreviated as DR hereafter.

We hypothesized that the model-based estimator (LGBM) would yield higher accuracy because the baseline model (DR) simply outputs the median of the training data. A one-tailed Wilcoxon signed-rank test was conducted on the per-subject mean *MAE* to test this directional hypothesis using a web-based tool[26] that implements the Wilcoxon test and Ryan's step-down correction. The same tool was used for subsequent testing.

As shown in Fig. 10, the LGBM model achieved a significantly lower *MAE* than DR for the Basic ($p = 0.001$, one-tailed, after Ryan's correction) and BoS ($p = 0.035$) features, but not for the NW features ($p = 0.079$). These results indicate that the LGBM clearly outperformed the median-based baseline when using the Basic and BoS features.

5.4 Combination of feature types

5.4.1 Introduction of integrated features

Integrated features were considered by combining different types of feature. Although no significant improvement was observed when using NW features alone, in Sect. 5.3, we imply that combining them with other feature types might yield higher results. Therefore, two types of integrated feature were introduced: one that integrates all three types, that is, the basic, NW, and BoS features, and the other that combines NW and BoS features, that is, symbol sequence features. They are represented as "int-sym" and "int-all," respectively. Regarding the NW and

Table 9
Relationship between N and feature computation time (s) ($\omega = 4$).

	$N = 5$	$N = 10$	$N = 15$	$N = 20$
NW features	0.10	0.12	0.15	0.15
BoS features	0.09	0.09	0.09	0.09

Table 10
Computation time for different BoS generation methods.

	Unigram	Bigram	Trigram
Time (s)	0.09	0.09	0.17

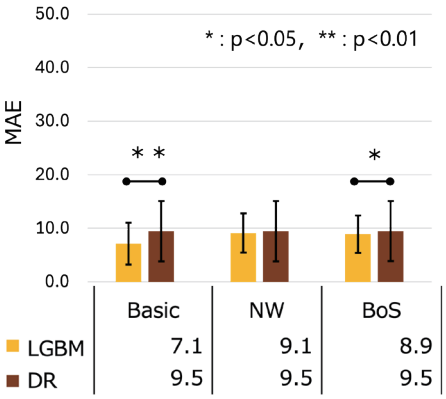

Fig. 10. (Color online) Comparison of estimation accuracy between different feature types ($\omega = 4$, $S = 120$, $N = 20$, and bigram for BoS).

BoS features included in each integrated feature, the optimal parameter settings identified in Sect. 5.2 ($\omega = 4$, $S = 120$, and $N = 20$) were used. BoS features were generated using bigram.

The estimation performances of the integrated features were compared with those of individual features. Specifically, the *MAE*, number of feature dimensions, and feature computation time were examined using data from the wrist sensor, which yielded the lowest estimation error. As an example of the results obtained for each of the three sensor positions, Table 11 lists the results using data obtained from the wrist. A two-tailed Wilcoxon signed-rank test was employed to assess significant differences in *MAE* among the feature types when using the LGBM model, with significance levels set at $p < 0.05$ and $p < 0.01$, as outlined in Table 12. In the pipeline for symbol sequence features (BoS and NW features), the calculation of symbolization features and their transformation into symbol sequences are common, as illustrated in Fig. 9. The subsequent computations of NW measures, such as in-degree and closeness centrality, and symbol occurrence probabilities are independent and thus can be performed concurrently. Although the computation time was measured sequentially, the tables report times assuming concurrent processing in which the longer of the two was adopted as the computation time for the symbol sequence features (int-sym). Similarly, for int-all, a longer interval between the basic and symbol sequence features was adopted.

Table 11
MAE, number of dimensions, and computation time for each feature.

	Basic	NW	BoS	int-sym	int-all
MAE	7.1	9.1	8.9	8.8	6.7
Number of dimensions	98	109	400	509	607
Computation time (s)	0.05	0.15	0.09	0.15	0.15

Table 12
Statistical test results for feature comparisons (values in parentheses indicate adjusted p-values).

	Basic	NW	BoS	int-sym	int-all
Basic	—	**(0.002)	**(0.002)	**(0.002)	**(0.004)
NW		—	(0.084)	**(0.004)	**(0.002)
BoS			—	(0.388)	**(0.002)
int-sym				—	**(0.002)
int-all					—

*: $p < 0.05$, **: $p < 0.01$.

The results indicate that int-all features performed significantly higher than the other types. However, the int-sym features demonstrated a higher accuracy than the NW and BoS features, although these differences were not statistically significant. These trends are listed in Tables 11 and 12. Additionally, the computation time for the int-all features was equivalent to that of the NW features, which had the longest computational time and required three times the feature computation time compared with the second-most accurate features, that is, basic features.

Beyond the feature-type-wise comparisons, the evaluation also considered differences across sensor positions. Specifically, *MAE* was compared for each sensor position using both LGBM and DR to estimate MWL. Since the DR model always outputs the target variable's median, we hypothesized that model-based estimation (LGBM) would yield higher accuracy. One- and two-tailed Wilcoxon signed-rank tests were used to compare models and sensor positions, respectively. Tables 13 and 14 show the test results and adjusted p-values.

As illustrated in Fig. 11 and Tables 13 and 14, the models using wrist data tended to yield lower MAEs across all feature types; however, significant differences from head data appeared only for the NW and int-sym features. For the most accurate int-all features, the models using the wrist and head data performed significantly higher than those using wiper data.

5.4.2 Importance of features

The importance of each feature was evaluated on the basis of the number of times each feature was used for node splitting during the training of the decision trees that constituted the LGBM. The top 40 features, along with their normalized importance, are shown in Fig. 12. The figure shows that the basic features tend to dominate the top rankings compared with the BoS and NW features; however, several NW features also appear in the top ranks with relatively high importance. This implies that the NW features may contribute meaningfully to the MWL estimation. Figure 13 shows the average feature importance for each feature group. Evidently, basic features account for the largest proportion, followed by NW features. In contrast, BoS

Table 13
Statistical test results for model comparisons across feature types (values in parentheses indicate adjusted *p*-values).

		DR, wrist	DR, head	DR, wiper
Basic	LGBM, wrist	**(0.001)		
	LGBM, head		**(0.001)	
	LGBM, wiper			**(0.002)
NW	LGBM, wrist	(0.079)		
	LGBM, head		*(0.049)	
	LGBM, wiper			(0.347)
BoS	LGBM, wrist	*(0.036)		
	LGBM, head		*(0.034)	
	LGBM, wiper			(0.754)
Int-sym	LGBM, wrist	*(0.036)		
	LGBM, head		*(0.021)	
	LGBM, wiper			(0.438)
Int-all	LGBM, wrist	**(0.001)		
	LGBM, head		**(0.001)	
	LGBM, wiper			**(0.002)

*: $p < 0.05$, **: $p < 0.01$.

Table 14
Statistical test results for sensor-position comparisons (values in parentheses indicate adjusted *p*-values).

		wrist	head	wiper
Basic	wrist	—	*(0.034)	**(0.004)
	head		—	*(0.019)
	wiper			—
NW	wrist	—	*(0.049)	(0.060)
	head		—	(0.638)
	wiper			—
BoS	wrist	—	**(0.006)	*(0.023)
	head		—	(0.937)
	wiper			—
Int-sym	wrist	—	*(0.012)	*(0.019)
	head		—	(1.000)
	wiper			—
Int-all	wrist	—	(0.117)	**(0.002)
	head		—	**(0.006)
	wiper			—

*: $p < 0.05$, **: $p < 0.01$.

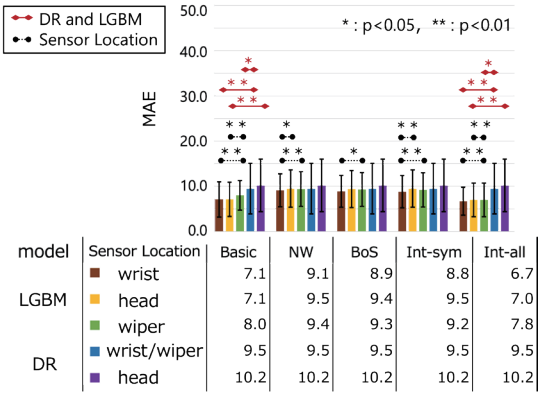

Fig. 11. (Color online) Comparison of integrated feature performance across sensor sites.

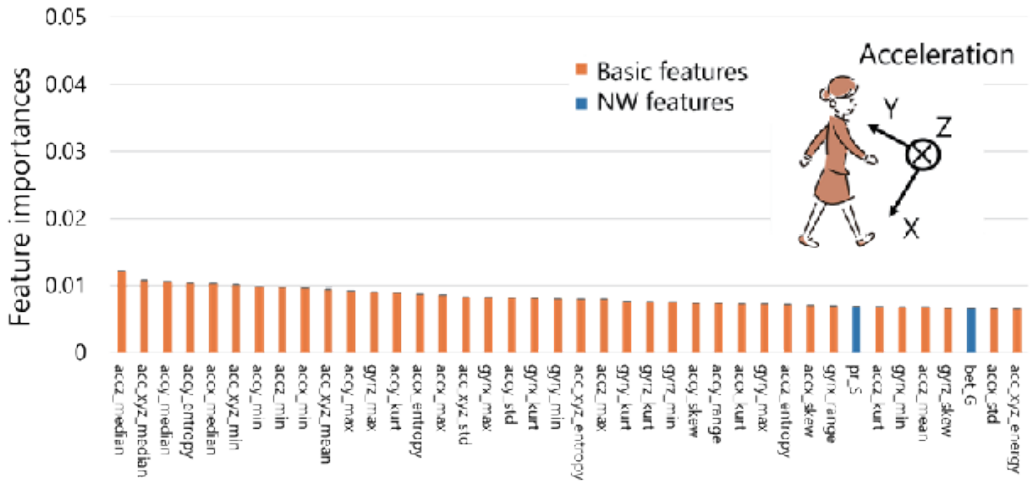

Fig. 12. (Color online) Normalized feature importance of top 40 features.

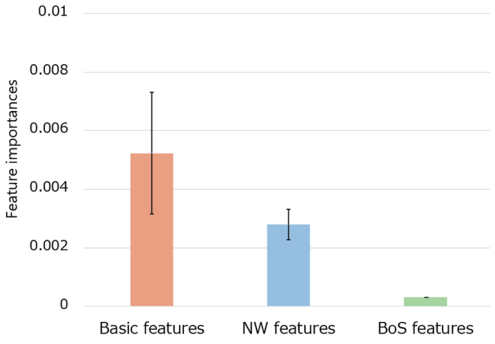

Fig. 13. (Color online) Percentage of total importance accounted for by each feature type.

features demonstrated lower average importance and reduced variability, suggesting their comparatively limited role in MWL estimation.

6. Discussion

6.1 Impact of parameters in symbol sequence features

As described in Sect. 5.2.1, smaller values of ω and larger values of N generally lead to improved estimation error. A smaller ω increases the number of symbol segments, S, within a single window, which enriches the information available for feature generation. This may decrease the error in MWL estimation. Similarly, a larger value of N increases the diversity of symbols within a window, resulting in more expressive symbol sequences.

In addition, BoS features based on bigrams and trigrams tend to outperform unigrams in terms of estimation errors. Although the difference is not statistically significant, as summarized

in Table 10, bigrams required less computation time than trigrams. This suggests that bigrams are more practical for BoS feature calculations.

6.2 Feature selection by usage scenarios

Two primary application scenarios were considered for MWL estimation. The first scenario involves analyzing long-term MWL changes for purposes such as healthcare, where daily data can be processed offline (e.g., at night), which makes immediacy unnecessary and minimizes estimation errors. The second scenario requires real-time use, such as providing alerts to a person who is cleaning, to prevent human error, where low latency is critical, even with some tolerance for errors. For the accuracy-focused scenario, the int-all feature set, that is, basic, NW, and BoS features, is recommended because it consistently achieved the highest performance; when a 0.15 s delay is acceptable, int-all is optimal. In the real-time scenario, the basic features are more practical, which increases *MAE* by only 0.4 while reducing the computation time to one-third.

Battery constraints must also be considered in wearable devices such as smartphones and smartwatches. Because int-all requires three types of feature calculation, it incurs the highest energy cost. Therefore, basic features are more feasible options under time or battery limitations.

6.3 Selecting sensor position

Sensor position is a crucial factor in the MWL estimation system design. As shown in Fig. 10, the wrist-mounted sensor achieved the lowest *MAE* for the three positions. With int-all features, the sensor on the wrist yielded *MAE* improvements of 0.3 over the head and 1.1 over the wiper. Although the difference between the wrist and the head was not statistically significant, the wrist consistently exhibited higher performance.

However, in real-world scenarios, smartwatches are typically worn on the nondominant wrist. In this study, the sensor was placed on the dominant hand holding the wiper, which raises concerns about whether comparable accuracy can be achieved using the nondominant hand. To mitigate this issue, alternative sensor positions, such as the head and wiper, should be considered. This can be achieved using earphone-embedded or appliance-integrated sensors. This results in more natural and unobtrusive sensing. As shown in Fig. 10, the head-mounted sensor produced the highest accuracy among the nonwrist positions with int-all features. If a decrease of 0.3 *MAE* compared with the wrist is acceptable, head-mounted sensing offers a practical and user-friendly alternative.

6.4 Feature importance

As illustrated in Fig. 10, the estimation accuracy improves with int-all features compared with using only the basic features. This suggests that either the NW or BoS features included in the integrated feature set may complement the basic features by capturing information that would otherwise be missed. As shown in Fig. 11, although most top-ranked features belong to

the basic feature type, two NW features are included among the top 40. This suggests that the NW features contribute meaningfully to the MWL estimation. Further analysis of Fig. 12 reveals that the NW features exhibit the second-highest average importance, following basic features. These findings indicate that symbol sequence features, which capture long-term activity transitions, enhance MWL estimation accuracy. The NW features appear to encode temporal structures that basic features alone cannot represent.

Additionally, basic features such as the median, minimum, and maximum values of acceleration contribute significantly to the accuracy. This can be attributed to the force applied by the wrist during floor cleaning, which likely reflects the MWL of the user. Uemura *et al.* reported that primary tasks tend to become automated under high-workload conditions. [9] Floor cleaning under high MWL can lead to decreased force applied to the wiper by reflecting more automated movement. Conversely, under low-workload conditions, actions may be more deliberate, involving clearer and stronger wrist motions. Thus, it is hypothesized that variations in the wrist force associated with MWL fluctuations play a critical role in the estimation.

6.5 Limitations

Despite these promising results, this study has several limitations. As discussed in Sect. 6.3, the participants in the data collection experiment were instructed to hold the wiper in their dominant hand with the sensor worn on the same wrist. Accordingly, the wrist data may reflect movement patterns specific to floor cleaning using the dominant hand. This raises concerns regarding the generalizability of the estimation accuracy when sensors are worn on a nondominant wrist, such as in typical smartwatch usage scenarios. In addition, the optimal parameter values were determined under the assumption that a user-specific regression model can be trained for each person, which may differ from a scenario in which a generic model is trained for individuals who were not present in the training dataset. Improving the generalization capabilities of the model is essential to achieve performance sufficient for practical applications. In the leave-one-subject-out cross-validation of wrist-worn sensor data, the lowest *MAE*s were 19.5, 18.6, and 19.0 for the basic, NW, and BoS features, respectively. These results suggest that a substantial performance gap remained between the user-specific and generalized models, which was likely due to strong individual differences. In future research, this gap should be narrowed by incorporating concepts such as transfer learning or domain adaptation. Furthermore, the reported feature computation times do not account for additional latencies caused by data collection and regression processing in actual system operation. Consequently, the actual end-to-end latency from sensing to estimation may be longer than that reported in this study.

7. Conclusion

We presented MWL estimation during floor cleaning as a case study of MWL estimation using our proposed method based on activity information. In addition to the conventional statistical features from inertial sensor signals, two types of feature derived from symbol sequences via vector quantization were defined. By carefully selecting the parameters of the

sensor positions and features, a minimum *MAE* of 6.7 was obtained, which demonstrates its feasibility. Regarding feature selection, the results indicated that employing all the integrated features yielded the lowest *MAE*. Conversely, using only basic features offers a practical advantage by significantly reducing computational cost while maintaining acceptable accuracy. Regarding the sensor position, the data obtained from the wrist achieved the lowest *MAE*; however, the results from head-mounted sensors imply the potential for realizing practical MWL estimation systems through natural sensing methods, such as inertial sensors integrated into earphones. Head data are also presumed to reflect torso movements, indicating the potential for extending this approach to other daily activities involving locomotion.

Future work will focus on estimating MWL during various tasks involving walking. We will investigate the generalizability of the estimation models using the data collected during walking, as well as the potential of adapting specialized models for walking to specific activities, such as floor cleaning using domain adaptation techniques.

References

1 M. Causse, F. Dehais, P.-O. Faaland, and F. Cauchard: Proc. 5th Int. Conf. Research in Air Transportation (2012) 2012. https://www.academia.edu/download/52869196/Causse_2012_ICRAT2012.pdf
2 J. Paxion, E. Galy, and C. Berthelon: Front. Psychol. **5** (2014) 1344. https://doi.org/10.3389/fpsyg.2014.01344
3 B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah: Proc. 2008 IEEE Int. Conf. Acoustics, Speech and Signal Processing (IEEE, 2008) 2041–2044. https://doi.org/10.1109/ICASSP.2008.4518041
4 L. Longo, C. D. Wickens, G. Hancock, and P. A. Hancock: Front. Psychol. **13** (2022) 883321. https://doi.org/10.3389/fpsyg.2022.883321
5 K. A. Brookhuis and D. de Waard: Accid. Anal. Prev. **42** (2010) 898. https://doi.org/10.1016/j.aap.2009.06.001
6 J.-P. Chaput and A. Tremblay: Obes. Rev. **11** (2010) 548. https://doi.org/10.1111/j.1467-789x.2010.00730.x
7 M. Cropley, L. W. Rydstedt, and D. Andersen: J. Epidemiol. Community Health **74** (2020) 919. https://doi.org/10.1136/jech-2019-213367
8 E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey: Proc. 12th ACM Int. Conf. Ubiquitous Computing (ACM, 2012) 301–310. https://doi.org/10.1145/1864349.1864395
9 Y. Uemura, Y. Kajiwara, and H. Shimakawa: Proc. 2016 Int. Conf. Computational Science and Computational Intelligence (IEEE, 2016) 1164–1167. https://doi.org/10.1109/CSCI.2016.0220
10 M. Shidahara, A. Tsuji, and K. Fujinami: Proc. 2024 IEEE 13th Global Conf. Consumer Electronics (IEEE, 2024) 144–145. https://doi.org/10.1109/GCCE62371.2024.10760683
11 T. Kosch: arXiv:2010.07703v2 (2020). https://doi.org/10.48550/arXiv.2010.07703
12 S. Mach, P. Storozynski, J. Halama, and J. F. Krems: Appl. Ergon. **105** (2022) 103855. https://doi.org/10.1016/j.apergo.2022.103855
13 S. Shimizu, T. Ito, Y. Yin, S. Arakawa, O. Sawada, and I. Aoyagi: Transactions of JSAE **50** (2019) 505 (in Japanese). https://doi.org/10.11351/jsaeronbun.50.505
14 S. Tanaka, A. Tsuji, and K. Fujinami: Int. J. Activity and Behavior Computing **1** (2024) 1. https://doi.org/10.60401/ijabc.10
15 S. G. Hart and L. E. Staveland: Adv. Psych. **52** (1988) 139. https://doi.org/10.1016/S0166-4115(08)62386-9
16 J. M. Noyes and D. P. J. Bruneau: Ergonomics **50** (2007) 514. https://doi.org/10.1080/00140130701235232
17 D. Conway, I. Dick, Z. Li, Y. Wang, and F. Chen: Proc. 14th IFIP TC13 Int. Conf. Human-Computer Interaction, P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler, Eds. (Springer, Berlin, Heidelberg, 2013) Part IV, pp. 659–666. https://doi.org/10.1007/978-3-642-40498-6_58
18 P. Chua, P. Sasikumar, Y. Weerasinghe, and S. Nanayakkara: arXiv:2409.12921v1 (2024). https://doi.org/10.48550/arXiv.2409.12921
19 F. Chen, N. Ruiz, E. Choi, J. Epps, M. A. Khawaja, R. Taib, B. Yin, and Y. Wang: ACM Trans. Interact. Intell. Syst. **2** (2013) 1. https://doi.org/10.1145/2395123.2395127
20 ATR-Promotions: https://www.atr-p.com/products/TSND121_151.html (Accessed October 2025).
21 F. Kawsar, C. Min, A. Mathur, and A. Montanari: IEEE Pervasive Comput. **17** (2018) 83. https://doi.org/10.1109/MPRV.2018.03367740

22 Y. Zhang, R. Jin, and Z.-H. Zhou: Int. J. Mach. Learn. Cyb. **1** (2010) 43. https://doi.org/10.1007/S13042-010-0001-0
23 S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang: Phys. Rep. **424** (2006) 175. https://doi.org/10.1016/j.physrep.2005.10.009
24 S. Brin and L. Page: Computer Network and ISDN Systems **30** (1998) 107. https://doi.org/10.1016/S0169-7552(98)00110-X
25 ELEGOO: https://tinyurl.com/2yo4qkee (Accessed August 2025).
26 M. Sugaya: https://stats.m-sugaya.jp/ (Accessed October 2025).