# Estimating Work Engagement with Wrist-Worn Heart Rate Sensors

Haruki Harashima*, Yutaka Arakawa*, Shigemi Ishida†, Yugo Nakamura*

* *Kyushu University*, Fukuoka, Japan, Email: {harashima.haruki, yutaka, yugo}@arakawa-lab.com.
* *Future University Hakodate*, Hokkaido, Japan, Email: ish@fun.ac.jp.

***Abstract*—This study aims to estimate the work engagement (WE) of office workers using biological data related to their daily activities obtained from wearable devices, with the goal of providing appropriate mental and physical health support to each individual. We collected daily heart rate data from wearable devices worn by 60 office workers in five Japanese companies for 2–3 weeks. Their daily WE was measured using the state-of the-art utrecht work engagement scale (UWES) questionnaire. We performed two types of analysis on the collected data using machine learning methods. First, the classification of binary WE levels (high or low), which showed a leave-one-person-out (LOPO) cross-validation F1 value of 0.522. Second, we classified whether WE decreased compared to the previous day, which showed a LOPO cross-validation F1 value of 0.663.**

***Index Terms*—Wearable Computing, Occupational Health, Machine Learning , Work Engagement, Heart Rate**

## I. Introduction

Recently, there has been a growing interest in using data related to human psychology, physiology, and the surrounding environment to improve the performance and comfort of workers. In particular, in Japan, the decline in the working-age population has led to a shortage of human resources, which in turn has decreased employee motivation and job satisfaction. The challenge for companies is to create workplaces where employees can work enthusiastically, especially in the COVID-19 era [1].

Work engagement (WE) is often used to quantify and measure work-related psychological states, such as positivity and fulfillment. Schaufuli et al. defined WE as "a positive and fulfilling work-related state of mind, characterized by vigor, enthusiasm, and immersion," which is considered the opposite of burnout, a state in which employees lose enthusiasm for work because of physical and mental fatigue [2]. Additionally, Harter et al. reported a correlation between employees' WE and labor productivity [3]. It is thought that increasing WE has excellent benefits not only for employees but also for organizations. In Japan, an increasing number of companies have introduced tools such as Motivation Cloud* and Wevox† to manage employees' WE through responses to online questionnaires.

*https://www.motivation-cloud.com/
†https://get.wevox.io/



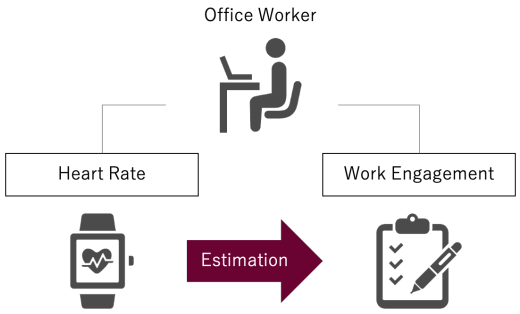


Fig. 1: Overview of our proposed method

Generally, questionnaires such as the utrecht work engagement scale (UWES) assess WE. However, such questionnaires are qualitative and subjective, which may lead to errors in the assessment of WE. Furthermore, the questionnaire is usually only answered once a month, which makes constant measurement difficult in terms of time and cost.

Against this background, our study aims to monitor the daily physical and mental state of workers using easily available wearable devices to support performance improvement and health management. We investigated the use of machine learning methods in conjunction with heart rate as a biometric to estimate daily WE.

Heart rate, especially heart rate variability (HRV), is directly influenced by autonomic nervous system activity and has previously been used for mental health analysis. For example, Huang et al. attempted to detect mental fatigue using HRV information obtained from electrocardiographs [4]. They used machine learning to classify binarized mental fatigue states and obtained an average accuracy of 75.5%. Additionally, Coutts et al. used HRV information obtained from a wrist-worn wearable device to estimate mental health states, such as depression, positivity, and anxiety [5]. This study shows that these states can be estimated with high accuracy using deep learning. However, their experiment was conducted on students and did not analyze work-related mental health conditions. These studies classified mental health status into two groups: high and low. In our study, we classify them in the same way. Additionally, we conducted a second experiment to estimate the daily variation in WE, as it is considered a practical

978-4-907626-48-8/21/$31.00 ©2021 IPSJ

approach for daily mental and physical health support.

Baethge et al. analyzed the relationship between HRV data obtained from civil servants and WE measured using questionnaires [6]. They concluded that there is a relationship between sympathetic activation and WE. Although they assessed the relationship between HRV and WE, the idea of predicting WE status has not been well tested. Therefore, in this study, we investigate whether the WE of office workers can be estimated using only features related to heart rate.

In this study, heart rate data obtained using wearable devices were collected from 60 office workers over a 2–3-week period. Simultaneously, we administered the UWES occupational health questionnaire and calculated the daily WE scores. We performed two experiments using these data and built a binary classification machine learning model for each.

In Trial 1, we classified binary WE levels, and obtained a LOPO cross-validation F1 value of 0.522.

In Trial 2, we classified each subject's WE score into two categories: a decrease in WE compared to the previous day, and an increase or no change in WE compared to the previous day. We obtained a LOPO cross-validation F1 value of 0.663.

We can see from the confusion matrix results that the "no change or increase" state was classified with high accuracy. These results show that wearable sensors can monitor the daily WE of office workers to a certain extent.

## II. Related Work

### A. Questionnaire Estimation Using Sensors

Various questionnaires have been used to assess mental health. For example, the Spielberger state-trait anxiety inventory (STAI) is the most frequently used measure of anxious mood in applied psychological research. The perceived stress scale (PSS) measures perceived stress levels, including chronic stress in daily life, the stress caused by concerns about the future, and stress caused by current circumstances. On the other hand, the depression and anxiety mood scale (DAMS) is a questionnaire used to measure the degree of depressed, positive, and anxious moods of the responder.

There is an existing research on estimating the results of such questionnaires using several sensors and machine learning methods. Table I shows the various questionnaires on mental health used in past studies and the sensor values used to estimate their results.

### B. Predicting Mental Health Using HRV

Huang et al. investigated the possibility of detecting mental fatigue using an electrocardiogram (ECG) device [4]. In their study, 35 students were asked to take an exam. A 14-item questionnaire measured the mental fatigue state before and after the exam. Simultaneously, HRV indices were collected at 5-min intervals using an electrocardiograph. A binary classification machine learning model was constructed to estimate the mental fatigue state using eight HRV indices from the time and frequency domains. A maximum accuracy of 75.5% was determined using five-fold cross-validation, and the best-performing algorithm was K-nearest neighbor (KNN) with an area under the curve (AUC) of 0.74. The results suggest that HRV is effective in assessing the mental state.

As in Huang et al.'s method, the HRV index is usually calculated from the heart rate interval (RRI) measured using an electrocardiograph. However, recently, alternative ECG methods have been used. For example, Coutts et al. analyzed questionnaire-measured mental health status data using the time and frequency domain features in conjunction with other basic HRV measures obtained from wearable sensors worn on the wrists of 652 students [5]. PSS, STAI, and DAMS were used as indicators of mental health status. Statistical analysis showed a significant difference between HRV measurements in the binarized mental health status, implying that HRV measures effectively diagnose mental health. Using long short-term memory (LSTM), a deep recurrent neural network commonly used in time-series analysis, a model was constructed to classify each mental health indicator as a binary level: high or low. Classification accuracies of 83% and 73% were obtained using approximately 2,000 5-minute and 500,000 2-minute HRV datasets, respectively.

HRV measures were used to predict the mental state of students, but not to predict WE.

### C. Statistical Analysis-Related Research on WE and HRV

Baethge et al. measured the frequency-domain HRV index in 118 civil servants over five days and analyzed the relationship between sympathetic activation and WE [6]. WE was obtained using a questionnaire based on the Utrecht Work Engagement Scale (UWES). The authors used multilevel analysis to test the hypothesis that there is a positive relationship between WE and sympathetic activation when workers are at work, on holidays, or sleeping.

Their study suggests a link between WE and HRV, but no insights into WE estimation have been proposed. However, we aim to develop a method to directly support the mental health of office workers using wrist-worn wearable devices.

## III. Proposed Method

### A. Data Collection

In this study, we used a dataset collected from a sensing project involving office workers [14]. The data were collected from 60 office workers in five Japanese companies by measuring their physical and mental state during daily life over a period of two to three weeks. Additionally, the subjects were required to complete a questionnaire on static characteristics, such as gender and age. During the experiment, they wore the Fitbit Charge 3[‡], which was used to measure the heart rate. The subjects also answered a questionnaire about WE every morning at 9 am. The total number of WE questionnaires responded to was 569, and the total amount of collected heart rate data using the Fitbit was 1.18 GB (13,424,969 records).

[‡]https://www.fitbit.com/global/us/home

TABLE I: Sensors Used to Estimate the Questionnaire

| Questionnaire | Reference | Sensors/Data |
|---|---|---|
| DAMS | Coutts et al. [5] | HRV |
| | Fukuda et al. [7] | Sleep information |
| STAI | Coutts et al. [5] | HRV |
| | Mozos et al. [8] | HRV, Electrodermal activity, Microphone, Acceleration |
| PSS | Coutts et al. [5] | HRV |
| | Sano et al. [9] | Acceleration, Screen time, GPS, Skin conductance |
| PANAS (Positive and Negative Affect Schedule) | Muaremi et al. [10] | Heart rate, Audio, Acceleration, GPS, Contact frequency |
| SAM (Stress-Appraisal Measure) | Giakoumis et al. [11] | Acceleration, HRV, Dermal activity, Video |
| WHOQOL (Quality of life for international comparison) | Amemori et al. [12] | Acceleration, GPS, Electrodermal activity, Heart rate, Skin temperature |
| OLBI (Oldenburg Burnout Inventory) | Garcia-Ceja et al. [13] | Acceleration |
| UWES | - | - |

### *B. Data Preprocessing of Questionnaire Responses*

In this paper, WE was used as an occupational health index to assess the active engagement and energy of a subject at work.

In our study, subjects were asked to respond by assigning a score of 0-6 to each of the three UWES-based questions: "I feel energized when I work now," "I am passionate about my work now," and "I am absorbed in my work now." The sum of these, the WE score, ranges from 0 to 18 and is used in subsequent analyses. Data corresponding to subjects who forgot to wear their Fitbit, or that did not answer the questionnaire, along with data showing a constant WE score throughout the experiment, were removed.

In Trial 1, the WE scores were classified and labeled into two high and low levels based on the median of the entire dataset. In Trial 2, we labeled the data based on the difference between the current day's WE score and the previous day's WE score. The two labels were "decrease" and "no change or increase" compared to the previous day. The sizes of the final datasets used in our experiment linking heart rate data and WE labels in each experiment were 317 and 275, respectively. Figure 2 shows the distribution of the WE scores for each subject, as measured by the questionnaire. This shows that there is a large difference between the median values and dispersion. Figure 3 shows the distribution of daily changes in the WE scores for each subject. This suggests no large bias in the number of decreases and increases in each subject's WE score since the median values are gathered around 0 (no change from the previous day).

### *C. Feature Extraction*

The Fitbits measure heart rate at a sampling rate of 0.2 Hz. This study calculated several features, including HRV features, based on the heart rate data over the following two periods:

Period 1: During sleep
Period 2: For one hour before the questionnaire (8–9am)

TABLE II: Extracted Features from Heart Rate Data

| | Features |
|---|---|
| 1 | Average Heart Rate |
| 2 | Median Heart Rate |
| 3 | Standard Deviation of Heart Rate |
| 4 | Maximum Heart Rate |
| 5 | Minimum Heart Rate |
| 6 | Average RRI (meanNN) |
| 7 | Standard Deviation of RRI (SDNN) |
| 8 | Root Mean Square of the Difference Between Adjacent RRIs (RMSSD) |
| 9 | Variance of RRI |
| 10 | Length of the Minor Axis of the Poincaré Plot (SD1) |
| 11 | Length of the Major Axis of the Poincaré Plot (SD2) |

Coutts et al. analyzed the relationship between nighttime and daytime HRV data and mental health obtained using questionnaires [5]. They suggested that nighttime and daytime HRV data represent differences in several mental health conditions. Therefore, we consider using nighttime and daytime heart rate data in this study. For daytime data, only the data before the questionnaire response was used. Furthermore, Period 1 was divided into three sections: total sleep time, after sleep onset, and before waking.

HRV features are calculated from the RRI, which is the interval between heartbeats. However, the RRI was not provided by Fitbit Charge 3. Therefore, we created a time-domain HRV feature by calculating the RRI from the heart rate. The calculation method of the RRI is shown in Equation 1, and the 11 extracted features are listed in Table II. The Poincaré plot (or Lorenz plot) is a scatter plot of the relationship between each RRI and the preceding RRI. These plots are generally elliptical, and their minor and major axis lengths (*SD1* and *SD2*) are used to assess autonomic activity [15], [16].

$$RRI = \frac{60}{HeartRateData} \times 1000 \tag{1}$$

In Trial 1, the difference and the ratio between Period 1 and Period 2 for each feature were added as new features. In

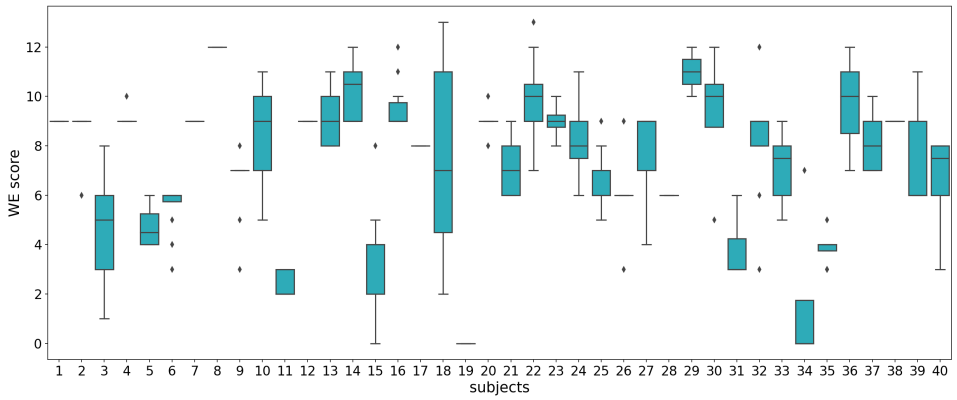


Fig. 2: Distribution of WE scores per subject

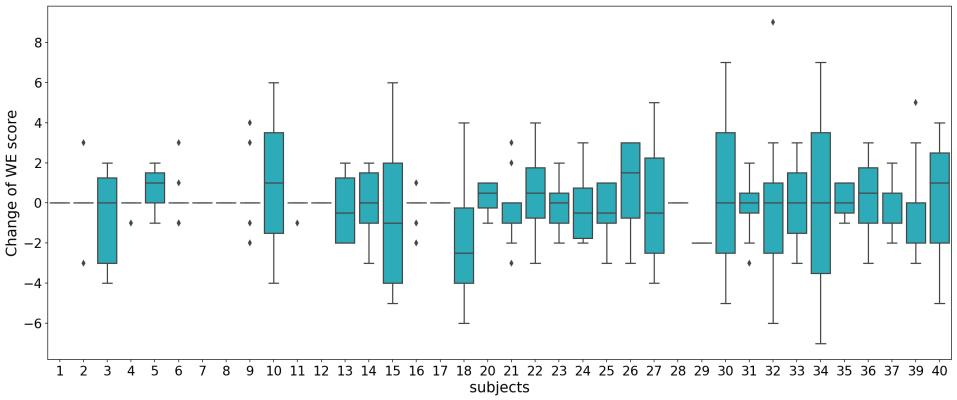


Fig. 3: Distribution of changes of WE score per subject

Trial 2, the ratios between the previous day's feature values and the current day's feature values were defined as features and used in addition to the features used in Trial 1. We also performed feature selection in each experiment by performing a statistical significance test between the two labels.

### *D. Model Building and Evaluation Method*

For both trials, we used a light gradient boosting machine (LightGBM) as the classification model. LightGBM is a type of gradient boosting decision tree, a popular machine learning algorithm in recent years owing to its high accuracy and computational speed [17]. Each model is trained to solve binary classification problems that determine the WE levels (high and low), or change in WE score (down, stay/up).

Models were evaluated using the Leave One Person Out (LOPO) cross-validation method. In the LOPO cross-validation method, the data of a single subject is used as the validation set, and the rest of the subjects' data were used as the training set. This method was repeated for all subjects to evaluate the generalization performance of the model among the subjects. Performance was measured using the accuracy, precision, recall, and F1 values.

## IV. Results and Discussion

### *A. Result of Trial 1*

Because of the LOPO cross-validation evaluation, the WE level was estimated with an accuracy of 0.576 and an F1 value of 0.522. The confusion matrix is shown in Fig. 4. Additionally, there no statistically significant difference in features between the two groups with high and low WE levels.

Box plots of the accuracy, precision, recall, and F1 values evaluated for each subject by LOPO cross-validation are shown in Fig. 5. From Fig. 5, the inter-subject variance of all indices was large, and the interquartile range of the accuracy and F1 value was 0.60 and 0.54, respectively.

Table III ranks the top ten features by average feature importance for classification using the LightGBM model, which mostly includes features related to sleep and features created for Trial 1. Because the heart rate is more stable during sleep than during the daytime, it is considered the baseline value for each subject. Ratio of the heart rate value

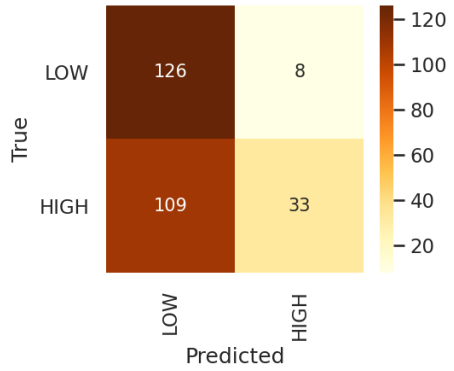


Fig. 4: Confusion matrix of Trial 1

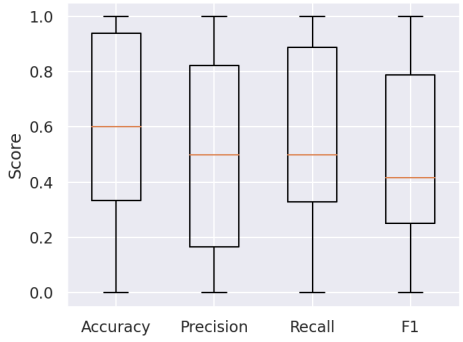


Fig. 5: Trial 1 box plots

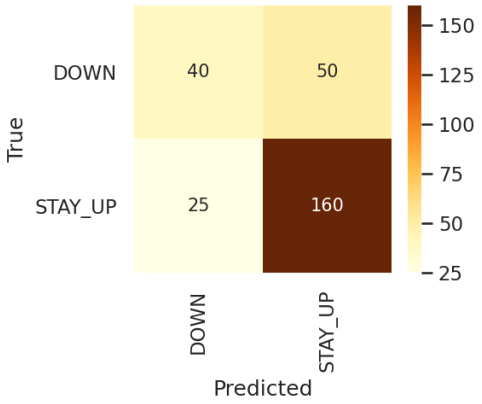


Fig. 6: Confusion matrix of Trial 2

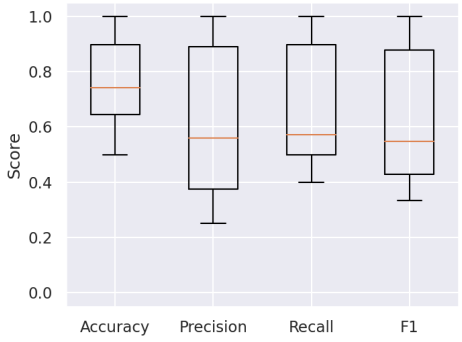


Fig. 7: Trial 2 box plots

during work to the baseline value is considered an effective feature for estimating WE. Among the ratio-related features, the *RMSSD* was the most important. The *RMSSD* is considered to have a strong relationship with the parasympathetic nervous system [18]. Thus, the difference in WE levels may be caused by the parasympathetic nervous system activity.

The confusion matrix shows that low-level WE detection is more accurate than high-level WE detection. There was also a large dispersion of F1 values among the subjects, with an interquartile range of 0.54. These results may be because we did not consider the possibility that the WE score norms differ between subjects. As shown in Fig. 2, the median and variance of the WE scores differ considerably between individuals. However, in this experiment, the WE-level classification was based on the median of the whole data and did not consider each subject's subjective interpretation of the questionnaire. We believe that setting and personalizing WE-level standards for each user can improve the performance of our method. However, this would necessitate the availability of additional data for each subject.

TABLE III: Feature Importance for Trial 1 Estimation

| Features | Remarks | Importance |
|---|---|---|
| Maximum Heart Rate | Period 1 | 0.044 |
| RMSSD | Ratio of two periods | 0.033 |
| Minimum Heart Rate | Ratio of two periods | 0.033 |
| Average Heart Rate | Period 2 | 0.029 |
| Standard Deviation of Heart Rate | Ratio of two periods | 0.029 |
| Minimum Heart Rate | Before waking | 0.029 |
| SD2 | After falling asleep | 0.029 |
| Average Heart Rate | Ratio of two periods | 0.028 |
| Median Heart Rate | Ratio of two periods | 0.027 |
| Maximum Heart Rate | Ratio of two periods | 0.024 |

### *B. Result of Trial 2*

The LOPO cross-validation results showed an accuracy of 0.709 and an F1 value of 0.597 when estimating the daily variation in WE score.

A statistical significance test was performed for all features that showed changes in WE scores between the two groups (decrease and other than decrease compared to the previous day), with 12 features found to be significantly different. The LOPO cross-validation results using these 12 features showed an accuracy and an F1 value of 0.727 and 0.663, respectively. The features with significant differences, along with their respective importance for LightGBM classification and $p$-values, are shown in Table IV. The corresponding confusion matrix is shown in Fig. 6. Among the features

TABLE IV: Feature Importance and P-value for Trial 2 Estimation

| Features | Remarks | p-value[1] | Importance |
|---|---|---|---|
| Standard Deviation of Heart Rate | Period 1 | 0.047 | 0.097 |
| Average Heart Rate | Period 2 | 0.007 | 0.068 |
| Median Heart Rate | Period 2 | 0.011 | 0.051 |
| Standard deviation of Heart Rate | Period 2 | 0.033 | 0.151 |
| meanNN | Period 2 | 0.013 | 0.130 |
| Average Heart Rate | Ratio of two periods | 0.015 | 0.095 |
| Maximum Heart Rate | Ratio of two periods | 0.049 | 0.062 |
| Standard Deviation of Heart Rate | Ratio of two periods | 0.006 | 0.079 |
| meanNN | Ratio of two periods | 0.034 | 0.096 |
| SDNN | Ratio of two periods | 0.029 | 0.000 |
| Variance of RRI | Ratio of two periods | 0.029 | 0.072 |
| SD2 | Ratio of two periods | 0.026 | 0.101 |

[1] Mann–Whitney U test

created for Trial 2 (i.e., features created using information from the previous day), the *standard deviation of heart rate* and *average heart rate* and *meanNN* showed a high level of importance. In particular, the *standard deviation of heart rate* has three types: Period 1, Period 2, and their ratio, and can be considered an effective feature for predicting changes in WE score. Fig. 7 shows the variations in the evaluation results among the subjects.

The confusion matrix shows that the model is better at detecting when the WE score increases or does not change than when it decreases. Moreover, the interquartile range of the accuracy and F1 values for each subject was 0.25 and 0.45. Compared to the evaluation results of Trial 1, the variability of the interquartile range of each evaluation index in Trial 2 was small, suggesting that the method for estimating the relative variability of WE scores is an approach with better generalization performance than the method for estimating absolute WE levels.

Previous studies to estimate mental health have not focused on daily intra-individual changes. However, we believe that estimating daily changes in mental health can contribute to the early detection and prevention of mental health problems and deterioration. In this experiment, we found that it is possible to detect changes in WE using features created as a ratio between the current and previous day's heart rate values

## V. Conclusion

In this study, we collected heart rate data from 60 office workers in five Japanese companies using wearable devices, along with their responses to the UWES questionnaire over a 2–3 week period. In Trial 1, a classification of the binary level of WE (high or low) was performed using a LightGBM classifier. The results of LOPO cross-validation showed an F1 value of 0.522. In Trial 2, we classified the change in WE score (a decrease compared to the previous day, or an increase or no change compared to the previous day). The result of the LOPO cross-validation showed an F1 value of 0.663.

We showed that it is possible to monitor the WE of office workers to some extent using wrist-worn heart rate sensors. In future work, we will aim to improve system performance by adding information such as daily sleep time and the number of steps walked. Furthermore, we will conduct experiments using a larger number of subjects to verify the generalization performance for practical applications.

## Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number JP18H03233, and the Cooperative Research Project Program of the Research Institute of Electrical Communication, Tohoku University.

## References

[1] A. Shimazu, A. Nakata, T. Nagata *et al.*, "Psychosocial impact of covid-19 for general workers," *Journal of occupational health*, vol. 62, no. 1, p. e12132, 2020.

[2] W. B. Schaufeli, I. M. Martinez, A. M. Pinto *et al.*, "Burnout and engagement in university students: A cross-national study," *Journal of cross-cultural psychology*, vol. 33, no. 5, pp. 464–481, 2002.

[3] "The relationship between engagement at work and organizational outcomes," https://employeeengagement.com/wp-content/uploads/2013/04/2012-Q12-Meta-Analysis-Research-Paper.pdf, 2012.

[4] S. Huang, J. Li, P. Zhang *et al.*, "Detection of mental fatigue state with wearable ecg devices," *International journal of medical informatics*, vol. 119, pp. 39–46, 2018.

[5] L. V. Coutts, D. Plans, A. W. Brown *et al.*, "Deep learning with wearable based heart rate variability for prediction of mental and general health," *Journal of Biomedical Informatics*, vol. 112, p. 103610, 2020.

[6] A. Baethge, N. M. Junker, and T. Rigotti, "Does work engagement physiologically deplete? results from a daily diary study," *Work & Stress*, pp. 1–18, 2020.

[7] S. Fukuda, Y. Matsuda, Y. Tani *et al.*, "Predicting depression and anxiety mood by wrist-worn sleep sensor," in *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 2020, pp. 1–6.

[8] O. M. Mozos, V. Sandulescu, S. Andrews *et al.*, "Stress detection using wearable physiological and sociometric sensors," *International journal of neural systems*, vol. 27, no. 02, p. 1650041, 2017.

[9] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 671–676.

[10] A. Muaremi, B. Arnrich, and G. Tröster, "Towards measuring stress with smartphones and wearable devices during workday and sleep," *BioNanoScience*, vol. 3, no. 2, pp. 172–183, 2013.

[11] D. Giakoumis, A. Drosou, P. Cipresso *et al.*, "Using activity-related behavioural features towards more effective automatic stress detection," *PloS one*, vol. 7, no. 9, p. e43571, 2012.

[12] Y. A. Chishu Amemori, Teruhiro Mizumoto and K. Yasumoto, "Simplified measurement method for hrqol based on whoqol-bref by smart devices," *Multimedia, Distributed, Cooperative, and Mobile Symposium*, vol. 2017, pp. 880–887, 2017.

[13] E. Garcia-Ceja, V. Osmani, and O. Mayora, "Automatic stress detection in working environments from smartphones' accelerometer data: a first step," *IEEE journal of biomedical and health informatics*, vol. 20, no. 4, pp. 1053–1060, 2015.

[14] Y. Tani, S. Fukuda, Y. Matsuda *et al.*, "Workersense: Mobile sensing platform for collecting physiological, mental, and environmental state of office workers," in *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 2020, pp. 1–6.

[15] L. Mourot, M. Bouhaddi, S. Perrey *et al.*, "Quantitative poincare plot analysis of heart rate variability: effect of endurance training," *European journal of applied physiology*, vol. 91, no. 1, pp. 79–87, 2004.

[16] J. Jeppesen, S. Beniczky, P. Johansen *et al.*, "Detection of epileptic seizures with a modified heart rate variability algorithm based on lorenz plot," *Seizure*, vol. 24, pp. 1–7, 2015.

[17] G. Ke, Q. Meng, T. Finley *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, pp. 3146–3154, 2017.

[18] F. Shaffer and J. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, vol. 5, p. 258, 2017.