IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 30 November 2024, accepted 16 December 2024, date of publication 25 December 2024, date of current version 8 January 2025.

Digital Object Identifier 10.1109/ACCESS.2024.3523108

RESEARCH ARTICLE

Detecting Subtle Signs of School Attendance Issues Using Smartphone-Based Sensing

VIKTOR ERDÉLYI[1], TERUHIRO MIZUMOTO[1,2], (Member, IEEE), YUICHIRO KITAI[1,3], DAIKI ISHIMARU[4], HIROYOSHI ADACHI[5], TERUO HIGASHINO[6], (Senior Member, IEEE), AND MANABU IKEDA[4]

[1]Graduate School of Information Science and Technology, Osaka University, Suita, Osaka 565-0871, Japan
[2]Department of Computer Science, Faculty of Innovative Information Science, Chiba Institute of Technology, Narashino, Chiba 275-0016, Japan
[3]Nippon Telegraph and Telephone West Corporation, Miyakojima-ku, Osaka 534-0024, Japan
[4]Department of Psychiatry, Graduate School of Medicine, Osaka University, Suita, Osaka 565-0871, Japan
[5]Health and Counseling Center, Osaka University, Toyonaka, Osaka 565-0871, Japan
[6]Department of Information and Computer Science, Kyoto Tachibana University, Kyoto, Kyoto 607-8175, Japan

Corresponding author: Teruhiro Mizumoto (mizumoto@ist.osaka-u.ac.jp)

This work was supported in part by the Innovation Platform for Society 5.0 from Japan Ministry of Education, Culture, Sports, Science and Technology under Grant S004541.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the Health and Counseling Center, Osaka University, under Approval No. 15, approved on September 2020, and the Ethics Committee of the Graduate School of Information Science and Technology, Osaka University, under Approval No. 20-42, approved in December 2020.

ABSTRACT In recent years, school attendance issues among university students have been increasing, which can lead to repeating courses, dropping out of school, or even social withdrawal. Despite the existence of counseling services, students often delay help-seeking, which can cause symptoms to worsen and make support more difficult. Thus, it is essential to identify at-risk students early and encourage them to seek help. A realistic approach must minimize the burden on students, rely only on devices they already own, and operate correctly even for students who are less engaged or prone to social withdrawal. While several techniques have been proposed to estimate individual indicators, they fail to address one of these requirements due to requiring additional devices or requiring user attention and interaction. In this paper, we propose an unobtrusive screening method for detecting subtle signs of school attendance issues in university students. We develop a smartphone app to collect sensor data and collect ground truth information using questionnaires for 1) sleep problems; and 2) decreased student engagement. We collect data from 58 university students for about 10 months, and build estimation models for the above indicators. Our evaluation shows that the estimation models are sufficiently accurate in flagging problematic cases. The indicators can then be used to notify at-risk students and medical practitioners, enabling timely intervention. This screening is not intended to replace traditional face-to-face medical examinations, but rather to selectively flag at-risk students and connect them with medical experts.

INDEX TERMS School attendance issues, sleep state estimation, subjective sleep quality estimation, study engagement estimation, smartphone sensing.

I. INTRODUCTION

In recent years, class attendance issues have been on the rise among university students. In particular, the long-term presence of (1) sleep problems [1] and (2) reduced study engagement [2] often leads to school attendance issues. These issues have caused many students to fail and/or repeat courses or to drop out of university entirely. To deal with this problem, many universities take mental health measures such as offering psychological counseling by clinical psychologists and psychiatrists. However, students often fail to recognize their own attendance problems and delay seeking help such as counseling, which can exacerbate their symptoms and make treatment more difficult. Therefore, early detection

The associate editor coordinating the review of this manuscript and approving it for publication was Alba Amato.

of students showing subtle signs of attendance issues, and encouraging them to get help, is important.

Recently, methods of monitoring a person's daily life and estimating their physical and mental state by using IoT devices, smartphones, and wearable devices have been proposed. These methods aim to estimate sleep habits [3], [4], [5] and various human psychological states such as study engagement [6], [7], [8], emotion [9], [10], depression [11], [12], stress [13], [14].

Moreover, several methods aim to monitor academic life-log data including sleep, academic performance (GPA), stress [15], [16], and emotion using smartphone and/or wearable devices. However, the applicability of methods using wearable and IoT devices is limited among students due to low device ownership rates. Additionally, existing methods either require extra devices (such as wearable and IoT devices) or require user attention and interaction, which may be difficult for students with low engagement and/or a tendency toward social withdrawal.

When designing a method for early detection, we need to meet some requirements to make large-scale adoption realistic. First, the method should not disrupt students' daily lives or require additional device purchases. It should also support students who may not actively seek help. Additionally, the system must be constructed using correct medical knowledge and correct ground truth data. Finally, since the onset of school attendance issues is often gradual, long-term and automated observations are crucial.

Our goal is to develop an unobtrusive screening method that can identify students at risk of developing school attendance issues. Using sensor data from smartphones and ground truth data from questionnaires, we aim to build accurate machine learning models to predict indicators related to school attendance issues. To minimize the burden on students, we design our method to require smartphones only (which are widely adopted), and operate in the background without user intervention.

We select sensors that are commonly available on standard smartphones. Specifically, our method collects Wi-Fi scan data, acceleration, gyroscope and magnetic field measurements, battery information, screen state, ambient light intensity, noise levels, and proximity information. With this data, we train machine learning models to estimate the responses to the questionnaires that measure indicators related to sleep and engagement. Regarding sleep, our primary objective is to detect a decline in overall sleep quality using *only* smartphone sensor data, *excluding* any external sensors or devices. Thus, despite the abundance of factors associated with sleep problems, in this paper, we focus only on those that can be estimated from smartphone sensor data. Specifically, we estimate sleep state (awake or asleep) and subjective sleep quality. Regarding study engagement, we estimate three standard engagement indicators (vigor, dedication, and absorption).

Our estimation models can enable meaningful applications, such as automatically giving feedback to students about their lifestyle, and allow students to present the data to healthcare professionals. Healthcare professionals can use this data to offer personalized advice, suggest interventions, or gather more information through questionnaires. Our proposed screening method is not intended to replace traditional face-to-face medical examinations, but instead to provide insights and encourage students to make changes or seek help before symptoms worsen.

Our contributions are as follows:

- In cooperation with the Osaka University Health Center and the Department of Psychiatry Graduate School of Medicine, we design and build a sensing-based student support platform for estimating the above indicators using medium- to long-term smartphone life log data only. Our system works with off-the-shelf Android smartphones and operates entirely in the application layer, without any operating system modifications.
- We evaluate our estimation techniques using 58 undergraduate and graduate students over the course of 10 months. Our evaluation shows that our models achieve either high or at least sufficient accuracy for the purposes of flagging problematic cases.

The rest of the paper is organized as follows. Section II describes related work on estimating psychological state using life log data, discusses existing methods related to estimating sleep and engagement, and clarifies this paper's position. Section III describes the overall design of our student support system. Section IV defines the specific indicators, explains their necessity, and describes our method for estimating them. Section V presents an experimental evaluation of our system. Finally, Section VI concludes the paper and describes future work.

II. RELATED WORK

In this section, we describe related work on sleep estimation and psychological state estimation, including research on engagement.

A. SLEEP ESTIMATION

Sleep estimation using wearable and/or mobile devices has been widely studied.

Borazio et al. [17] measured sleep rhythm and quality in based on heart beat data collected from a smartwatch. While devices such as smartwatches have higher accuracy, it has been reported that the quality of sleep deteriorates due to their invasiveness [18]. Additionally, despite their potential utility, smartwatch-based monitoring methods remain constrained in their applicability due to the relatively low adoption rates of these devices among undergraduate students.

There is also work on smartphone-based sleep estimation. For example, Wahl et al. [19] estimates users' fall-asleep and wake-up time from various smartphone sensors such as acceleration, illuminance, and sound. To improve the accuracy of the obtained results, they use additional information on each individual's sleep rhythm. They can estimate the fall-asleep time with an absolute error of 42 to 48 minutes, and the wake-up time with an absolute error of 42 to 57 minutes. Chen et al. [4] constructed a model that estimates sleep time using a regression model. Hao et al. [20] constructed a system that evaluates sleep quality by detecting events closely related to sleep quality, such as body movements and snoring, using only the smartphones' built-in microphones.

B. ESTIMATION OF PSYCHOLOGICAL STATES USING LIFE LOGS

In recent years, psychological state estimation using life logs (records of daily activities) has been gaining attention.

MIMOSYS [21] is a smartphone app that recognizes users' mental health condition from voice samples. To avoid reporting bias, the system detects the involuntary reaction of the vocal cords. Using features such as voice energy, the system can distinguish "patient", including major depression, cerebral infarction, etc., from "healthy person" with over 90% accuracy. MIMOSYS requires active measurement of 13 voice samples to collect one data point, making it very time-consuming. This type of approach, which relies heavily on user-initiated actions, could pose a significant barrier to continued use, especially for users who are struggling with their mental health and may already be experiencing difficulty with motivation and engagement.

Huynh et al. [22] estimate happiness using life logs for the purposes of detecting depression in students. They estimate happiness with 72% accuracy based on smartphone data (phone and SMS history, screen on/off status, GPS), and smartwatch data (such as heartbeat and electrodermal activity) for one month. Another work by Ito et al. [23] uses the sensors and usage history of smartphones to predict the responses of an anxiety questionnaire. They create a prediction model from data for several days for an individual, and predict the anxiety level for the next day. Using up to 7 days of data, they achieved an accuracy of about 80% in estimating anxiety. However, these systems do not comprehensively estimate study engagement.

There are also several works dealing with the issue of estimating engagement specifically [6], [7], [22], [24]. However, such systems tend to rely on invasive sensors (e.g., wristband, external depth camera, EDA, PPG, or even EEG) to perform the estimation. One system, TaskyApp [25], estimates task engagement (task difficulty) using only smartphones, but it mainly focuses on estimating the difficulty of mental tasks performed in an office setting. Automatic sensing is limited (triggered when a change in the user's context is detected), making TaskyApp unsuitable for longer-term tracking of study engagement in daily life.

Despite these advances, existing methods often struggle with university students' needs - particularly regarding device adoption, user burden, and the need for unobtrusive monitoring. Many systems rely heavily on user interaction, limiting their effectiveness for disengaged students. In the next section, we propose a smartphone-based system that operates continuously in the background to detect subtle signs of school attendance issues.

III. METHODOLOGY: SYSTEM DESIGN

In this section, we first define the requirements of the student support system and how we address each of them. Then, we present the design of the proposed system.

A. REQUIREMENTS AND APPROACH

As described in Section I, an effective school support system must satisfy a set of requirements and challenges. Below, we present our approach to each of these.

1) LOW BURDEN ON STUDENTS

To make the system easy to deploy, we should not use devices with low penetration rates (such as wearables [26], [27]), and we should not require additional devices per person. Therefore, we use only smartphones for data collection, which most university students already own [28]. Additionally, since we want to collect data from regular users' daily-use smartphones, we cannot expect them to apply modifications to their smartphones (e.g., custom OS). Therefore, we implement our data collection application entirely in the application layer as a regular Android app. To support passive students who do not actively engage in help-seeking behavior, we design our app so that it collects data automatically in the background.

2) CORRECT MEDICAL KNOWLEDGE AND GROUND TRUTH

To obtain domain knowledge, we worked together with the Osaka University Health Center and the Department of Psychiatry, Graduate School of Medicine and developed questionnaires to compute indicators related to school attendance issues. In traditional face-to-face interviews, biases may arise from students' subjective self-evaluation and memory limitations, as they're asked to recall events from a month ago. To tackle this issue, we use daily questionnaires to gather ground truth. While subjectivity remains, we argue that asking about recent events reduces the recall bias when compared to traditional face-to-face interviews.

B. SYSTEM OVERVIEW

Figure 1 shows the overall concept of the proposed student support system, which essentially represents a feedback loop. First, users provide smartphone sensor data and answer questionnaires related to indicators of school attendance issues. Using the collected data, we build and train estimation models to estimate the questionnaire responses based on the sensor data. After training, the system can estimate the relevant indicators from the sensor data without relying on questionnaires, enabling unobtrusive monitoring in the background. These indicators can then be evaluated by psychiatrists, who can give personalized feedback to users, or request them to fill out additional questionnaires.

The actual technical components of our system are shown on Figure 2. The system consists of a smartphone application that collects sensor data and questionnaire responses, a public cloud server, a local server hosted on our campus, a data analyzer application, and a PC used by doctors to view relevant metrics about the data. (In this paper, the feedback mechanism is not yet implemented.) In the following, we discuss each component of our system in detail.

1) SMARTPHONE APPLICATION

The smartphone application aggregates sensor data and questionnaire responses from participants through a unified interface, minimizing user burden. Figure 3 illustrates the application's Japanese-language interface, with English translations provided via speech bubbles to maintain accessibility for non-Japanese readers while preserving the original design.

To encourage active participation in data collection, we designed a point-based incentive system that gives points depending on the amount of data provided. Additionally, we changed the design of the home screen according to the daily questionnaire's response status (Figure 3, B). We also created a workflow for viewing the history of one's own rewards. This screen (Figure 3, D) shows the amount of points and the types of data collected for each day. We also have a simple weekly ranking system that creates friendly competition among users.

Our system supports the collection of the following sensor data:

a: WI-FI SCANS

One scan result consists of a list of (MAC address, received signal strength) pairs of nearby access points. Our app periodically initiates Wi-Fi scans, but the OS may delay or ignore some requests to save power.

b: ACCELERATION, GYROSCOPE AND MAGNETIC FIELD

3-axis acceleration, gyroscope and magnetic field data. Collection is continuous, but the data is periodically aggregated before storage.

c: BATTERY STATUS

The possible values are "charging", "discharging", "full", "not charging", "unknown". Data is collected when the charging status changes.

d: BATTERY LEVEL

The remaining charge of the battery, as a percentage. Data is collected when the remaining battery level changes by at least 1%.

e: SCREEN STATE

A binary value expressing whether the screen is on or off. Data is collected when the screen turns on or off.

f: ILLUMINANCE

Ambient light intensity. Data collection is constant, but the data is periodically aggregated before storage.

g: NOISE LEVEL

Root-mean-square values of audio data from the smartphone microphone. While technically we are recording potentially sensitive audio data, to alleviate privacy concerns, we only upload statistical features of the data to the server. Data is collected in 5-second bursts, one burst per minute.

h: PROXIMITY

Measures whether something is in proximity of the sensor. The value is an integer between 0 and 5 (closest proximity being 5). Data collection is continuous, but the data is periodically aggregated before storage.

Our system also supports conducting questionnaires. Specifically, it supports single-choice and multi-choice questions, a compound question about sleep times (going-to-bed time, fall-asleep time, and wake-up time), and star-based rating questions (e.g., subjective sleep quality).

Questionnaire data is sent immediately upon response, but sensor data is temporarily cached on the device in encrypted form, and periodically uploaded to the cloud server in the background. Both sensor data and questionnaire responses are encrypted using the public key of the local server before being uploaded to the cloud server. (Additionally, this communication happens using HTTPS, which provides an additional layer of encryption during transit between nodes.)

2) CLOUD SERVER

The cloud server acts as an intermediary between the smartphone app and the local server. It accepts upload requests from the smartphone app and download requests from the local server. This server needs to be publicly accessible so that users can access it regardless of their location. However, while we keep some plaintext metadata in anonymized form on this server, for security reasons, all sensitive user data is stored encrypted on this server, and only temporarily (see below).

3) CAMPUS LOCAL SERVER

The local server stores raw data securely and can only be accessed from the university network. Since it can't be contacted directly by the cloud server, it regularly checks for new data. When found, it downloads and decrypts the data, then deletes the encrypted version from the cloud server. The security of the raw data is protected by measures such as access control rules, full-disk encryption, making the data by accessible only to a limited number of people (such as doctors), and HTTPS communication.

4) DATA ANALYZER

The data analysis operates via a client application that interfaces with the local server's API endpoints within the campus

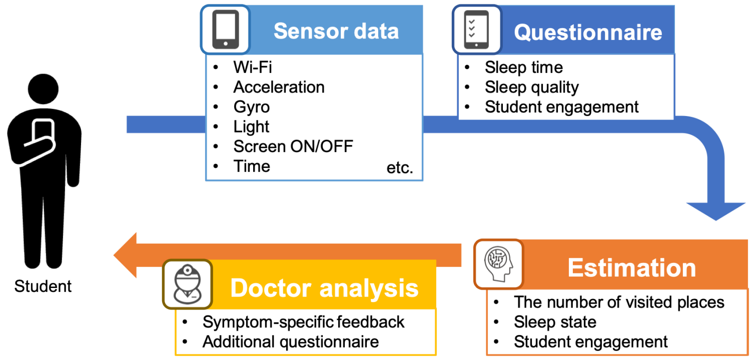


FIGURE 1. Overall concept of the proposed system.

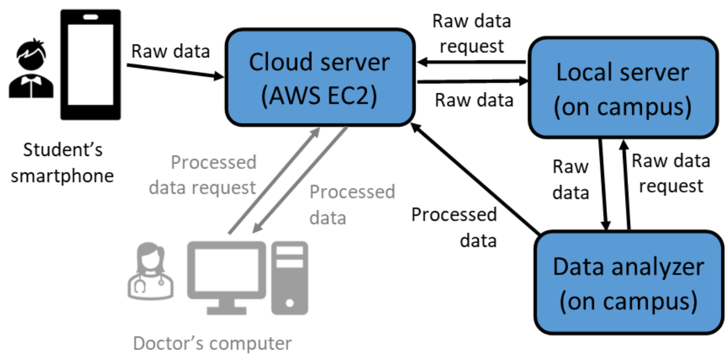


FIGURE 2. Technical components of the proposed system.

network. Access to raw data is secured through TLS client certificate authentication. The analyzer synthesizes sensor data and questionnaire responses with domain expertise from medical professionals to evaluate indicators of student life patterns, academic engagement, and school attendance behaviors. Future implementations could facilitate encrypted feedback delivery (including AI-derived insights, physician recommendations, and follow-up assessments) to users' devices via public key cryptography.

In summary, our system design ensures low user burden and accurate data collection through coordinated operation of smartphone, cloud, and local components. We now detail the estimation methods used to derive meaningful indicators from sensor data and questionnaire responses, focusing on sleep problems and decreased study engagement.

IV. METHODOLOGY: ESTIMATION METHODS

In this section, we describe our method for estimating indicators of school attendance issues. The high-level flow of the method is as follows. We first give a definition of the metrics we want to estimate, namely sleep problems and decrease of student engagement. Then, we design a questionnaire to obtain the corresponding ground truth data. Then, we design a method to extract features from smartphone sensor data. After these steps, we are ready to collect data (questionnaires and sensor data) and train classifier models to perform the actual estimation (Section V).

A. ESTIMATING SLEEP PROBLEMS

Based on sensor data collected by smartphones, we construct a classification model that can estimate (1) sleep state and (2) subjective sleep quality. We define these as follows:

Sleep state: a binary value that represents whether the user is asleep or awake during any given time slot. This definition was inspired by the sleep diaries used in traditional medical counseling. The sleep diary is a subjective evaluation of sleep rhythm, typically represented as a table, with the date along the Y axis, and the times of the day along the X axis (see Figure 4). Students indicate the times spent asleep by coloring the corresponding cells.

Subjective sleep quality: the user's self-evaluation of their sleep quality on an integer scale from 1 to 5 (1 = worst, 5 = best).

To obtain ground truth information, we ask users to answer a questionnaire about their sleep each day after they wake up. Table 1 shows the content of this questionnaire, which includes questions about the timing of sleep, and a five-star subjective sleep quality rating. To minimize user burden, we avoid asking about nighttime wake-ups. While this subjective self-evaluation is imperfect, it is regularly

(a) Home screen (number of daily questionnaire responses this week, current ranking, total points)

(b) Point information for a given week (weekly ranking, number of points, and the types of data provided)

(c) Evening questionnaire asking about study engagement

(d) Settings screen with checkboxes for Wi-Fi, GPS, other sensors, and noise level

FIGURE 3. **Screenshots from the smartphone application (original Japanese user interface with English translations in speech bubbles).**

used in medical counseling, and the relationship between subjective sleep quality and mental health has been explored extensively in the literature [29], [30], [31].

1) SLEEP STATE ESTIMATION

a: FEATURE EXTRACTION

To estimate sleep state, we first need to extract features from the raw data. In this study, we use the features shown in Table 2. We process the sensor data periodically and extract statistical information from them to be used as features.

For acceleration and gyro sensor data, features include (1) statistics of the 3D vectors from the sensors and (2) differences between consecutive values of these statistics in the time series. The latter "diffXX" features can be useful in determining whether a smartphone has moved.

For light and audio data, due to differences in the environment (location of the home, lighting during sleep), the relative differences in the data carries more useful information, so we standardize the data for each subject.

The "time" feature indicates the "time-slot of day" of the data collection. For example, if data is collected at 10-minute intervals, the day can be divided into 144 numbered time slots.

The "cluster" feature is a binary feature that indicates whether the user is at home or not. This feature is useful because we expect a large difference in sleep metrics between home and other places (see Section IV-C for more details).

TABLE 1. **English translation of the sleep questionnaire. (Questions were asked in Japanese.)**

Item	Answer format
At what time did you go to bed last night?	date/time
At what time did you fall asleep last night?	date/time
At what time did you wake up this morning?	date/time
How would you evaluate your sleep quality as a whole?	star rating (1 star = worst, 5 stars = best)

b: CLASSIFICATION

The next step is to classify each time slot.

Our sleep dataset is imbalanced, featuring more samples in the "awake" state compared to the "asleep" state due to the average sleep time of humans. Since this negatively imbalance affects the classification accuracy, we construct a Balanced Random Forest model, which can handle unbalanced data by adjusting the number of samples for each decision tree, and validate the model using Cross-Validation.

c: POSTPROCESSING

Since we estimate sleep state separately for each time slice, due to subtle inaccuracies in the classification model, spurious "asleep" time slots may appear in its output randomly during daytime activity. To improve the accuracy of the model, we try to correct such clear false positives after classification via smoothing. Figure 5 shows an overview of post-processing method.

We post-process the data in two ways. First, asleep states shorter than 1 hour are converted into the awake state. Second, awake states of 30 minutes or less between two asleep states are considered to be temporary awakenings, and we replace those with asleep states. In a previous experiment, we observed a slight improvement compared to the case without postprocessing, indicating that the postprocessing method is effective.

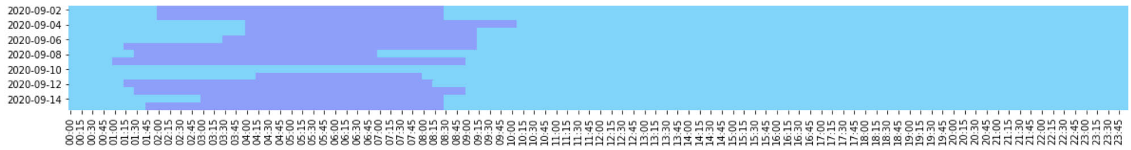


FIGURE 4. Example of sleep diary used in medical counseling.

TABLE 2. Features used for sleep state estimation.

Name	Detail	Name	Detail
accmax	maximum of acceleration	time	time slot
accmin	minimum	batterystatus	Charge/Non-Charge/Full
accavg	average	batterylevel	battery level
accstd	standard deviation(SD)	scstate	screen state ON/OFF
diffaccmax	difference of maximum	lightmax	maximum of illuminance
diffaccmin	difference of minimum	lightmin	minimum
diffaccavg	difference of average	lightavg	average
diffaccstd	difference of SD	lightstd	SD
gyromax	maximum of gyro	audiomax	maximum of noise volume
gyromin	minimum	audiomin	minimum
gyroavg	average	audioavg	average
gyrostd	SD	proximitymax	maximum of proximity
diffgyromax	difference of maximum	proximitymin	minimum
diffgyromin	difference of minimum	proximityavg	average
diffgyroavg	difference of average	proximitystd	SD
diffgyrostd	difference of SD	cluster	at home/not at home

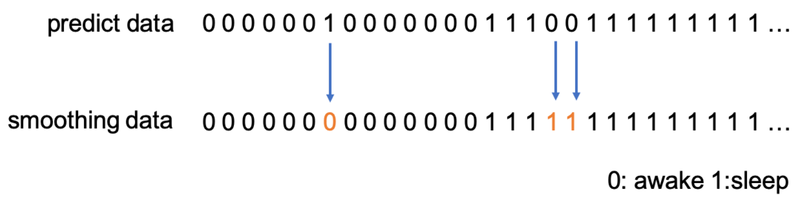


FIGURE 5. Example of our postprocessing (smoothing) method for sleep state estimation.

2) ESTIMATING SUBJECTIVE SLEEP QUALITY

The subjective sleep quality for a given night is a function of not just a single night but also of the environmental conditions and sleep patterns of previous days. Therefore, we use data from a certain period of time to estimate sleep quality. We aggregate the features shown in Table 2 per hour, and we estimate subjective sleep quality by using the feature data of a total of 31 hours (see Table 3 for the set of features).

TABLE 3. Features considered estimating subjective sleep quality.

Feature name	Detail
$SENSOR\{00-30\}$	Hourly statistics for each sensor
sleeptime	Sleep time
usetime	Usage time of smartphone
num_cluster	The number of visited places per day
weekday	weekday/holiday

Let t be a given user's wake-up time on a given day d. When extracting features for day d, we used the data in the time range from $t - 31$h to t. The reason for setting the time to be 31 hours is that we want to include the previous day's sleep in the input of the feature extraction. Assuming that the daily life rhythm is not significantly disturbed, we think this is reasonable.

The $SENSOR\{xx\}$ features come from Table 2. $SENSOR$ represents the sensor name, and xx represents the number of hours before the wake-up time t. For example, if the wake-up time is 09:00, accavg02 represents the average value of acceleration from 06:00-07:00.

Additionally, we use some features that do not have separate values for each hour, but rather one aggregate value:

- **sleeptime:** the total sleep duration within the time range $[t - 31\text{h}, t]$.
- **usetime:** the total smartphone usage time within $[t - 31\text{h}, t]$. Usage time is defined as the total time the smartphone screen is on, computed in 5-minute time slots. One 5-minute slot is considered 5 minutes of usage if the screen is turned on at all during that period, and 0 minutes otherwise.
- **num_cluster:** the total number of visited places within $[t-31\text{h}, t]$. For active individuals, this value is expected to be high. On the other hand, those who tend to isolate

themselves may exhibit lower values approaching 1. See Section IV-C for more details.

- **weekday:** a binary feature representing whether the time instant t is a holiday or a weekday; we added this feature because some people may get better quality of sleep on holidays.

The next step is to classify each time slot. Since the data used to estimate subjective sleep quality follows a pattern similar to the sleep state data, we use Balanced Random Forest for subjective sleep quality estimation as well.

B. ESTIMATING STUDY ENGAGEMENT

To obtain ground truth information, we ask users to fill out a 7-class questionnaire based on the Utrecht Work Engagement Scale (UWES) [32] every day in the evening before falling asleep, assessing the indicators of study engagement (vigor, dedication, and absorption). The specific items of this questionnaire are shown in Table 4, and the possible choices are explained in Table 5. Since our target audience was Japanese students, we used the Japanese version of the questionnaire in our study. As discussed in Section IV-A, our ground truth for study engagement relies on subjective evaluation.

TABLE 4. **Utrecht work engagement scale (UWES) [32].**

Indicators	Item	Choices
Vigor	When I'm doing my work as a student, I feel bursting with energy.	7 choices (0-6)
Dedication	I am enthusiastic about my studies.	7 choices (0-6)
Absorption	I am immersed in my work.	7 choices (0-6)

TABLE 5. **Choices of the Utrecht work engagement scale (UWES) [32].**

Point	Item name
0	Never
1	Almost never
2	Rarely
3	Sometimes
4	Often
5	Very often
6	Always

1) FEATURE EXTRACTION

There are several factors that affect study engagement. First, if the learning environment is quiet and calm, the concentration level may increase. Second, there are also temporal dependencies; for example, being busy and having learning difficulties during the previous day may affect the motivation of the next day. Third, it is also necessary to consider the influence of recovery time after learning. (Recovery is the free time after studying. It is generally spent on leisure activities, and is essential for sufficient study engagement on the next day.)

To capture such temporal dependencies (e.g., include the recovery time of the previous day), similarly to subjective sleep quality estimation, we aggregate the features shown in table 2 per hour. Then, we estimate engagement by using a total of 31 hours of feature data ($[t - 31\text{h}, t]$). However, unlike in sleep quality estimation, we used the bedtime as the reference time t.

2) CLASSIFICATION

To enhance battery life and user experience, we need to improve estimation accuracy and limit data collection. To do this, it is key to identify essential features for high accuracy. To obtain feature importance information and deal with imbalances in our questionnaire responses as mentioned in Section IV-A, we employ a Balanced Random Forest model and verify it using Cross-Validation.

C. COMPUTING FEATURES RELATED TO LOCATION CLUSTERS

In this section, we describe the "cluster" and "num_cluster" features used in our estimation models. First, the "cluster" feature refers to the ID of the user's location cluster at a given time. Subsequent visits to the same location will be identified as the same cluster without providing an absolute location. We only use this to determine whether the user is at home or not, and the "home" cluster is identified as the cluster that the user spends the most time at. Second, the "num_cluster" feature refers to the number of visited places. Changes in this number are closely linked to psychological factors like motivation and can be seen clearly among students with school attendance problems. When students lack motivation to go to class, they skip classes, resulting in less time spent at the university. Additionally, in cases such as social isolation or depression, they might not leave their homes at all.

We estimate the number of visited places from Wi-Fi access point (AP) scans (MAC address and RSSI). For each month and each user, we preprocess their respective data separately by pruning irrelevant data. For each scan (containing a list of APs with RSSI), we select the top five APs and create a binary vector that represents whether an AP was seen within that observation. For example, if the known set of access points is [a,b,c,d,e,f,g,h,i], and at a given time the observed access points are [c,b,d,a,f], then the corresponding binary vector will be [1,1,1,1,0,1,0,0,0].

Then, we use these binary feature vectors to perform Agglomerative Hierarchical Clustering (AHC), which involves iteratively merging "objects" (AP scan results) into clusters based on their dissimilarity. Since places near a certain location tend to have similar Wi-Fi RSSI signatures, we use the Euclidean distance between RSSI vectors as the dissimilarity metric. To combat overcounting (such as in cases where the user is traveling between places), we perform smoothing by keeping only the most observed cluster as the representative cluster within a given time interval, eliminating spurious clusters that have few observations.

To validate our Wi-Fi-based estimation method, we compare the results to ground truth data computed from GPS coordinates (GPS data is only used as a reference point to validate our method's accuracy). For each month, we use Vincenty's formulae [33] to calculate distances between GPS coordinates and cluster them using a threshold of 100m to account for indoor GPS error. The mean absolute error (MAE) is 1.017 and the root mean squared error (RMSE) is 1.761. The RMSE is higher than the MAE due to overestimation on certain days, presumably due to walking indoors with many Wi-Fi access points. However, the MAE indicates an error of approximately one point, suggesting that the method is useful.

V. EXPERIMENTAL EVALUATION

In this section, we evaluate the effectiveness of each model described in Section IV. First, we describe our application deployment, and then we evaluate our method's ability to estimate the relevant indicators, such as sleep and study engagement. Additionally, to reduce the battery consumption of the sensing system, we explore whether we can reduce the number of features collected for our estimations without losing too much accuracy.

In the below sections, we determine the optimal hyper-parameters of each estimation model by grid search and evaluate them via 10-fold cross-validation. The chosen hyper-parameters are shown in Table 6. Based on our empirical observations, utilizing only features above a significance threshold either enhances estimation accuracy or maintains performance parity, while reducing computational complexity. Therefore, we present results exclusively for this subset of high-importance features. We report the effectiveness of our estimation techniques via confusion matrices. These matrices show to what extent we can estimate the given mental state based on the training data (questionnaire responses). A higher estimation accuracy indicates a better ability to detect at-risk students and provide appropriate feedback about them.

TABLE 6. Hyperparameters obtained by grid search for each Random Forest estimation model (max_depth and n_estimators represent the max tree depth and the number of estimators, respectively).

Estimation model	Model variant	max_depth	n_estimators
Sleep status	binary	5	60
Subjective sleep quality	3-class	5	90
Subjective sleep quality	5-class	5	80
Vigor	3-class	5	100
Vigor	7-class	5	80
Dedication	3-class	5	60
Dedication	7-class	5	60
Absorption	3-class	5	100
Absorption	7-class	5	70

A. EXPERIMENTAL REAL-WORLD DEPLOYMENT

We performed a real-world application deployment to undergraduate and graduate students of Osaka University between April 24th, 2022 and February 28th, 2023 (about 10 months). To recruit participants, in April 2022, we displayed an advertisement within the campus management system of Osaka University called KOAN (Knowledge of Osaka University Academic Nucleus). The advertisement contained a link to our experiment website, which only allowed registrations from people who have a university-provided e-mail address. We received a total of 87 registrations from the website, out of which 58 users installed and activated our data collection app on their daily-use Android smartphone via the QR code that we e-mailed to them.

Table 7 shows some basic statistical information about the participants including age, gender, PSQI score, engagement and study stage. In this study, we did not collect information about their health status or academic performance.

TABLE 7. Participant information. (Avg=average, SD=standard deviation).

Metric	Value
Age	Avg: 21.24, SD: 2.32
Gender	40 male, 18 female
Study stage	45 Bachelor, 12 Master's, 1 Doctoral
PSQI score	Avg: 8.35, SD: 2.74
Study engagement score	Avg: 3.04, SD: 1.23

During the course of the study, participants answered the morning and evening questionnaires. In the background, our application collected sensor data automatically. This experiment was carried out with the approval of the Ethics Committee of the Health and Counseling Center, Osaka University (approval ID: 15, approved in Sep. 2020) and the Ethics Committee of the Graduate School of Information Science and Technology, Osaka University (approval ID: 20-42, approved in Dec. 2020). All participants provided their informed consent prior to enrollment in the study by reading the description of the study displayed on the sign-up page, and checking a box indicating their agreement. The sign-up page contained information such as the purpose of the research, the types of data we collect, how the data will be stored and analyzed (e.g., anonymization), compliance with relevant laws and the ethical guidelines related to human subjects research, an explanation of the ability to opt out or to selectively provide data at any time, a description of the rewards provided, and an option to agree to the secondary use of the collected data.

Data validity in our study was defined by the simultaneous presence of sensor readings and questionnaire responses (subjective ground truth). The dataset, aggregated across all participants, contained 2,255 days' worth of valid morning questionnaire data and 1,725 days' worth of valid evening questionnaire data. Model training was conducted exclusively using these valid records.

B. RESULTS AND ANALYSIS

1) SLEEP STATUS ESTIMATION

We performed our evaluation using 5-minute time slots. We estimated sleep state for each 5-minute time slot, and rounded the questionnaire responses (which were input at a resolution of 1 minute) to the nearest 5-minute mark. We computed each feature shown in Table 2 for each 5-minute interval, and we constructed a Balanced Random Forest model using these features that outputs an asleep/awake value for each time slot.

a: ESTIMATION RESULT

The confusion matrix of sleep status estimation and more detailed metrics are shown on Figure 6 and Table 8, respectively. The results show a recall of about 83% for 'awake (0)', 89.5% for 'asleep (1)', and an F1-score of 0.808 (macro average), while the precision of sleep estimation was about 61%. We think that the lower precision is due to the following factors. First, users were only able to input one sleep session per day, making it impossible to input daytime naps. In these cases, the system may correctly estimate an "asleep" state since the phone is not used, but the ground truth data will incorrectly indicate an "awake" state. Second, an awake user may have left their smartphone in a quiet and dark location for a period of time, causing the system to think that the user is sleeping. Finally, users' subjective evaluation of the start and end times of sleep sessions may be inaccurate.

The important features (defined as those with an importance value higher than 0.01) were, in decreasing order of importance, as follows: 'time', 'diffgyroavg', 'diffaccmax', 'gyrostd', 'accmax', 'audioavg', 'diffaccstd', 'lightavg', 'diffgyromax', 'lightmin', 'gyroavg', 'lightmax', 'diffgyrostd', indicating an understandable correlation with the time of day, movement, ambient lighting, and background noise level.

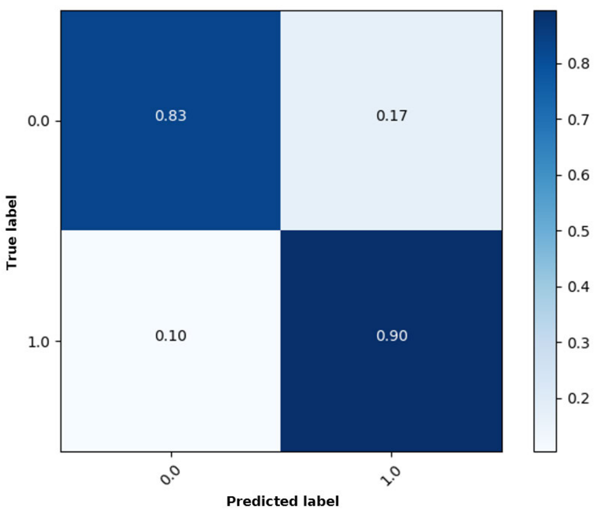


FIGURE 6. Confusion matrix of sleep state estimation. The label 0 means awake, and the label 1 means asleep.

TABLE 8. Evaluation metrics of sleep state estimation.

	precision	recall	f1-score	support
awake(0)	0.9640	0.8301	0.8921	401528
asleep(1)	0.6087	0.8951	0.7246	118518
accuracy			0.8449	520046
macro avg	0.7863	0.8626	0.8083	520046
weighted avg	0.8830	0.8449	0.8539	520046

2) SUBJECTIVE SLEEP QUALITY ESTIMATION

Based on the subjective sleep quality answers from the sleep questionnaire (Table 1), we built a Balanced Random Forest model to estimate the sleep quality rating using the features in Table 3. In the following, for the sake of simplicity, we may refer to subjective sleep quality as "sleep quality".

Since the number of features used for estimation is very large, we only enable features that have an importance of at least 0.01 in our preliminary evaluation. This value was chosen empirically. Figure 7 shows the important features for 5-class sleep quality estimation. The "Observation timing" value indicates the timestamp of the data, expressed as hours before the wake-up time. For example, if the observation timing is 04, it indicates that it is a feature of 4 hours before the wake-up time.

The results indicate that illuminance is a very important factor, especially "lightmin". Another important feature is "audiomin", an indicator of noise level. There are also some important features clustered around the observation timing of 20, indicating a connection between sleep quality and the activities during the previous daytime. These feature might be important for various reasons: (1) A bright environment may interfere with sleep, leading to a decrease in sleep time and sleep quality (2) High levels of daytime background noise indicate students were socially active and involved in extracurriculars, leading to deeper sleep and better overall sleep quality. (3) High levels of nighttime background noise may have a negative effect on sleep quality. Additionally, the aggregate (non-hourly) features "sleepTime", "num_cluster", and "use_time" were important. Sleep time is known to be an important factor in determining sleep quality. As for the number of clusters and smartphone use time, one possible reason for their importance could be that excessive activity or sensory overload could potentially impact sleep quality negatively, whereas moderate levels of exploration and stimulation might have a positive effect on sleep.

TABLE 9. Evaluation metrics of 5-class subjective sleep quality estimation. The labels represent the sleep quality rating (1 = worst, 5 = best sleep quality).

	precision	recall	f1-score	support
1	0.8203	0.8259	0.8231	293
2	0.5851	0.5494	0.5667	344
3	0.5070	0.4586	0.4816	556
4	0.6134	0.6248	0.6190	645
5	0.5702	0.6523	0.6085	417
accuracy			0.6035	2255
macro avg	0.6192	0.6222	0.6198	2255
weighted avg	0.6017	0.6035	0.6017	2255

The confusion matrix and the detailed results are shown on Figure 8 and Table 9, respectively. The table shows that the 1-star rating is estimated with a precision of 82%, but the precision of the other rating classes is low, leading to an overall F1-score of 0.62 (macro avg).

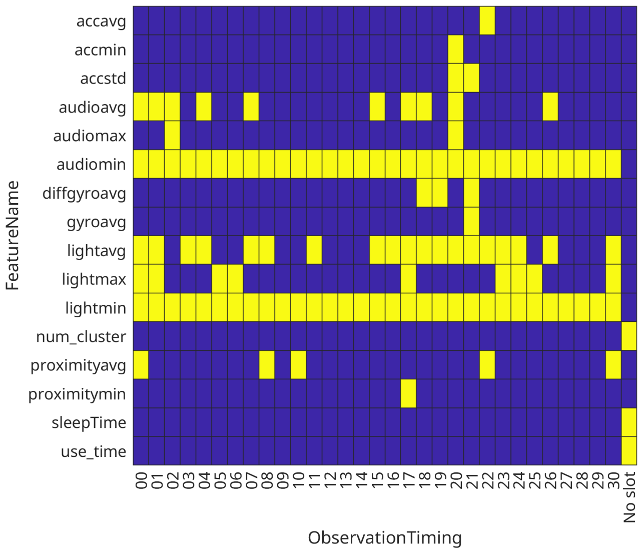


FIGURE 7. **Important features of 5-class subjective sleep quality estimation (Yellow = important, Dark blue = not important). Features without an observation timing are shown in the 'no slot' category.**

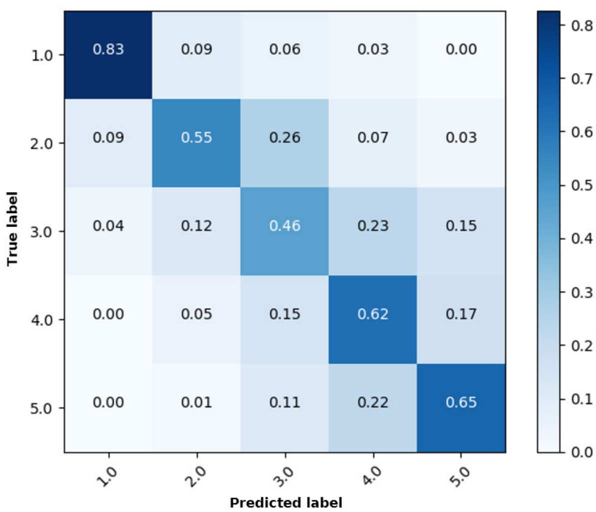


FIGURE 8. **Confusion matrix of 5-class subjective sleep quality estimation. The labels represent the sleep quality rating (1 = worst, 5 = best sleep quality).**

While the overall accuracy is low, the confusion matrix shows that most erroneous predictions are semantically close to the true answers (e.g., 2 stars instead of 1 or 4 stars instead of 5), and larger discrepancies were rare. This result shows that our method captures the quality of sleep well.

a: SIMPLIFIED ESTIMATION

For the purposes of supporting student life, the ability to estimate the overall tendency (is the situation good or bad) is more important than precisely estimating the quality of sleep. To evaluate our method in this aspect, we converted the 5-level sleep quality scale into a 3-level scale by combining the two highest ratings (4-5) into the label 3 and the two lowest ratings (1-2) into the label one. As before, we perform 3-class classification using only the important features. The important features are shown on Figure 9, which shows a similar pattern, i.e., a general importance of noise levels and ambient lighting, and similar timings for other important features.

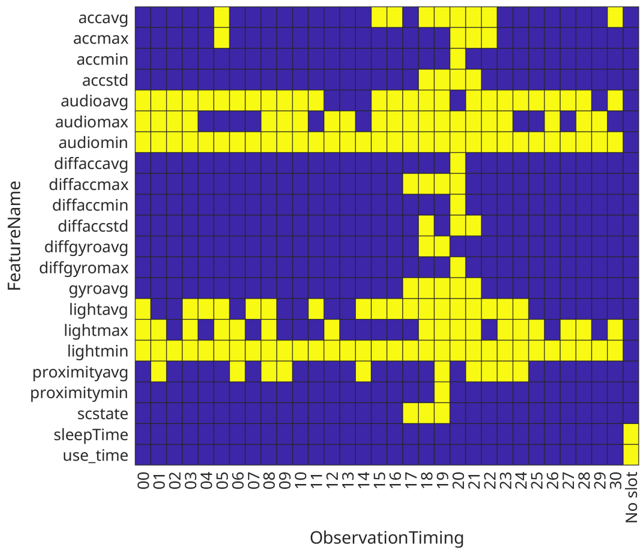


FIGURE 9. **Important features of 3-class subjective sleep quality estimation (Yellow = important, Dark blue = not important). Features without an observation timing are shown in the 'no slot' category.**

The results of 3-class estimation are shown in Figure 10 and Table 10. The confusion matrix shows that it is rare to confuse 0 (bad sleep) and 2 (good sleep), highlighting our method's usefulness for detecting a deterioration of sleep quality.

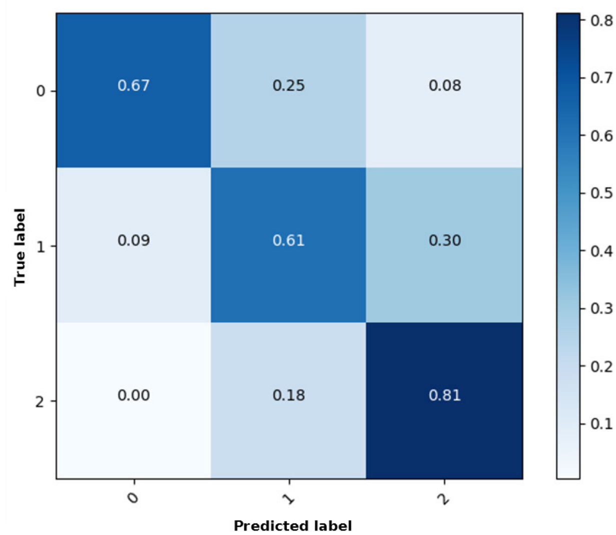


FIGURE 10. **Confusion matrix of 3-class subjective sleep quality estimation (important features). The labels 0, 1, and 2 represent bad, neutral and good sleep quality, respectively.**

TABLE 10. Evaluation metrics of 3-class subjective sleep quality estimation.

	precision	recall	f1-score	support
0	0.8908	0.6656	0.7619	637
1	0.4878	0.6115	0.5427	556
2	0.7967	0.8117	0.8041	1062
accuracy			0.7211	2255
macro avg	0.7251	0.6963	0.7029	2255
weighted avg	0.7471	0.7211	0.7277	2255

3) STUDY ENGAGEMENT ESTIMATION

Similarly to sleep quality estimation, the number of features is large, which may lead to a decrease in accuracy due to unnecessary features so we only enable the important features in our Balanced Random Forest model. In all engagement estimation models below, the cut-off threshold for important features was an importance value of 0.001. Also similarly to sleep quality estimation, our focus is on detecting trends and changes in study engagement, not on estimating the current state in the highest possible detail. Therefore, we also investigate a simplified version of the Utrecht Work Engagement Scale (UWES) [32] that contains only three classes (0-1, 2-4, 5-6). To distinguish between good and bad study engagement and detect its deterioration, it is sufficient to be able to distinguish the 5-6 (Positive) and 0-1 (Negative) classes.

Overall, noise levels and ambient lighting conditions are important features for estimating study engagement regardless of their observation timing. These environmental conditions can affect attention, concentration, and mood. For example, high noise levels can increase stress, leading to fatigue and decreased focus, while adequate lighting supports visual processing and comfort. Regardless of the time of day, these factors impact students' ability to learn, focus and stay engaged in their studies.

Another important feature was "num_cluster". A very high number of visited locations could be distracing and lead to decreased focus and productivity, ultimately impacting their ability to retain information and feel satisfied with their studies. Additionally, frequent transitions between locations can add up to stress and fatigue, further hindering effective learning. On the other hand, too few locations per day could suggest a lack of variety in their routine. A lack of diversity in environments and stimuli can lead to monotony and boredom, decreasing motivation and study engagement.

Finally, "use_time" as also important. If the smartphone usage time on a day is too high, it may indicate that the student is spending too much time on non-academic activities such as social media, gaming, or mindless scrolling, which can lead to procrastination and reduce their focus on studying. On the other hand, if the smartphone usage time is too low, it might imply that the student is not utilizing technology sufficiently to support their learning (e.g., by seeking out educational materials or exchanging information with their peers), leading to reduced motivation.

a: VIGOR ESTIMATION

The important features for vigor estimation are shown on Figure 11. In addition to the observations at the top of Section V-B3, there is a cluster of important features around an observation timing of 11-13, i.e., 11-13 hours before bedtime, which typically corresponds to various daytime activities.

The results of vigor estimation are shown on Figure 12 and Table 11. We were able to estimate vigor with an F1-score of 0.65 (macro avg). In cases where the estimation was wrong (typically in neutral answers), there were only minor discrepancies (the predicted class had similar scores to the true class).

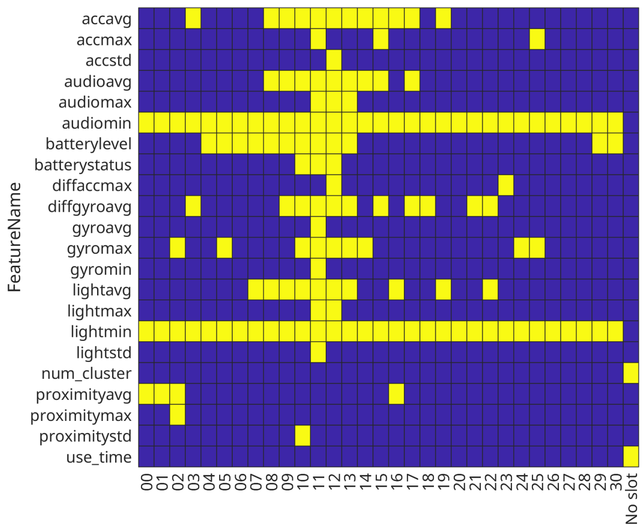


FIGURE 11. Important features of 7-class vigor estimation (Yellow = important, Dark blue = not important). Features without an observation timing are shown in the 'no slot' category.

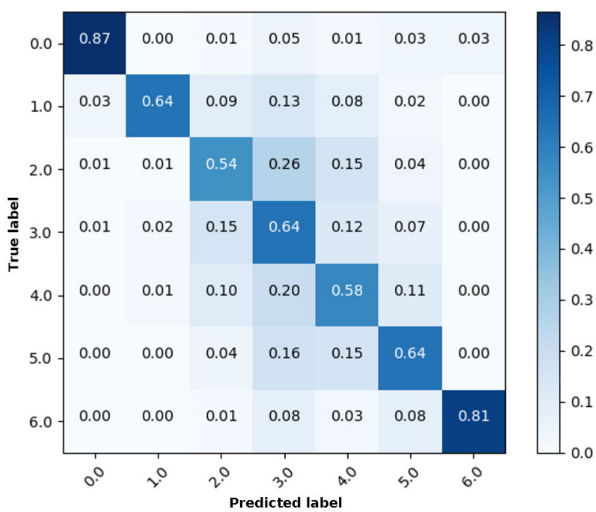


FIGURE 12. Confusion matrix of 7-class vigor estimation (0 = lowest vigor, 6 = highest vigor).

Regarding the simplified, 3-class estimation model (the mapping is described at the beginning of Section V-B3), the

TABLE 11. Evaluation metrics of 7-class vigor estimation.

	precision	recall	f1-score	support
0	0.9588	0.8651	0.9095	215
1	0.7703	0.6404	0.6994	89
2	0.3735	0.5439	0.4429	171
3	0.7674	0.6398	0.6978	758
4	0.5307	0.5775	0.5531	284
5	0.3598	0.6413	0.4609	92
6	0.9126	0.8103	0.8584	116
accuracy			0.6597	1725
macro avg	0.6676	0.6741	0.6603	1725
weighted avg	0.7014	0.6597	0.6734	1725

important features are shown on Figure 13, and the results are shown on Figure 14 and Table 12. While the precision of the class "Positive" was low, the macro avg F1-score is around 0.75.

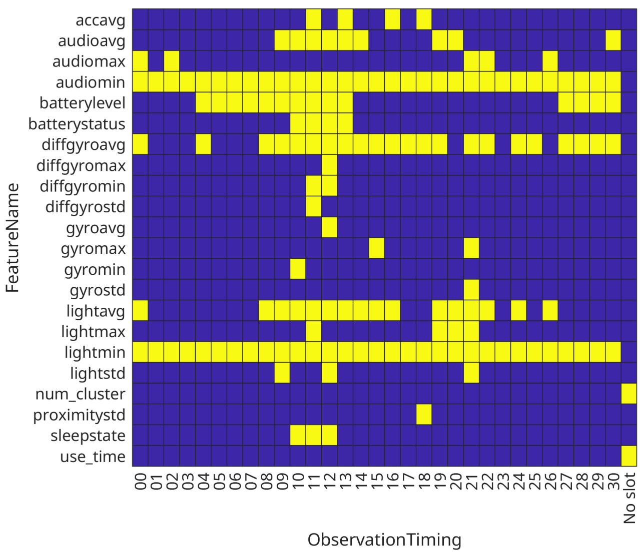


FIGURE 13. Important features of 3-class vigor estimation (Yellow = important, Dark blue = not important). Features without an observation timing are shown in the 'no slot' category.

TABLE 12. Evaluation metrics of 3-class vigor estimation.

	precision	recall	f1-score	support
0 (Negative)	0.8219	0.7895	0.8054	304
1 (Neutral)	0.8083	0.8030	0.8056	1213
2 (Positive)	0.6277	0.6455	0.6365	208
accuracy			0.7510	1725
macro avg	0.7527	0.7460	0.7492	1725
weighted avg	0.7527	0.7510	0.7518	1725

b: DEDICATION ESTIMATION

The important features of dedication estimation are shown on Figure 15. In addition to the observations at the top of Section V-B3, we can also see a cluster of important features around the observation timings of 10-12 representing daytime activities. Additionally, movement-related features like accelerometer and gyroscope seem to be more relevant

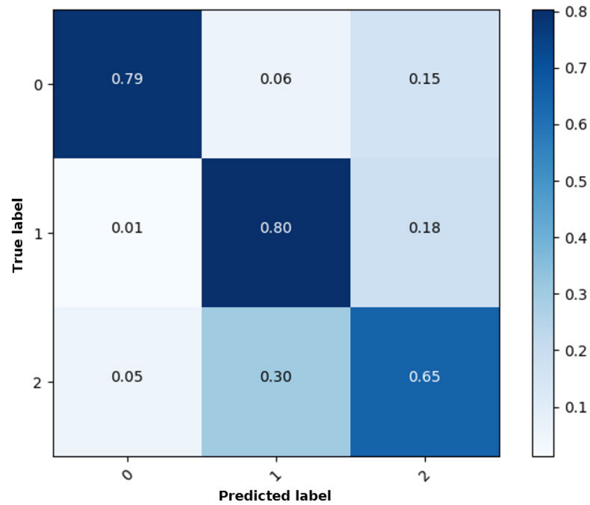


FIGURE 14. Confusion matrix of 3-class vigor estimation (0 = lowest vigor, 2 = highest vigor).

for dedication than in earlier sections, suggesting that physical activity during the day affects the dedication.

The results are shown on Figure 16 and Table 13. Our model achieved an overall macro-avg F1-score of 0.657, and has a low precision for class 5, presumably due to insufficient support. The confusion matrix suggests that, overall, incorrect estimations remain semantically close to the true label (within one point), except the relatively high number of cases where the model incorrectly outputs a neutral value (3). The value 3 happens to be the class with the largest support, suggesting an imbalanced dataset.

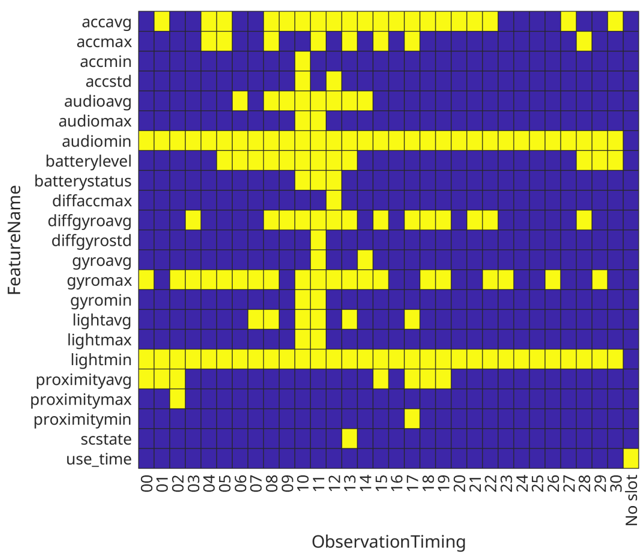


FIGURE 15. Important features of 7-class dedication estimation (Yellow = important, Dark blue = not important). Features without an observation timing are shown in the 'no slot' category.

We also evaluated our method with the simplified 3-class estimation model (with the mapping described at the

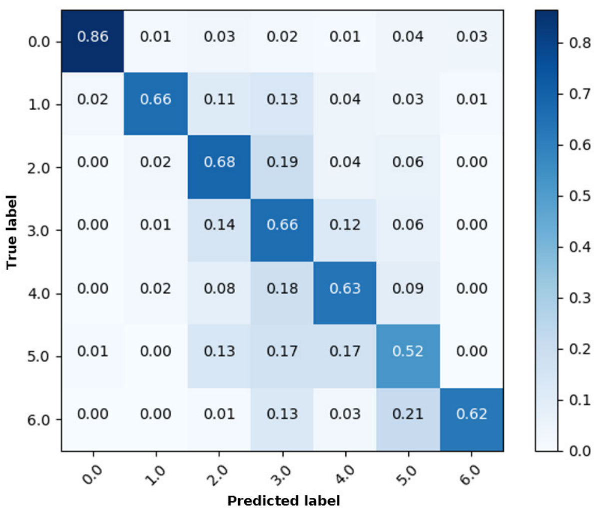


FIGURE 16. Confusion matrix of 7-class dedication estimation (0 = lowest dedication, 6 = highest dedication).

TABLE 13. Evaluation metrics of 7-class dedication estimation.

	precision	recall	f1-score	support
0	0.9795	0.8643	0.9183	221
1	0.7349	0.6559	0.6932	93
2	0.4847	0.6842	0.5675	209
3	0.7737	0.6577	0.7110	707
4	0.5771	0.6339	0.6041	254
5	0.2765	0.5222	0.3615	90
6	0.9216	0.6225	0.7431	151
accuracy			0.6736	1725
macro avg	0.6783	0.6630	0.6570	1725
weighted avg	0.7210	0.6736	0.6880	1725

beginning of Section V-B3). The results are shown on Figure 18 and Table 14. The model has an F1-score of 0.745 (macro avg). While it has some difficulty in distinguishing between "Neutral" and "Positive", it can estimate the (more important) negative cases with high accuracy.

TABLE 14. Evaluation metrics of 3-class dedication estimation.

	precision	recall	f1-score	support
0 (Negative)	0.8531	0.7771	0.8133	314
1 (Neutral)	0.7948	0.7778	0.7862	1170
2 (Positive)	0.6134	0.6572	0.6345	241
accuracy			0.7375	1725
macro avg	0.7538	0.7374	0.7447	1725
weighted avg	0.7426	0.7375	0.7395	1725

c: ABSORPTION ESTIMATION

The important features are shown on Figure 19. Similar to earlier sections, the influence of noise level and ambient lighting remains strong, and we can still observe a cluster of important features at the observation timings 10-13. This result indicates that the user's environment throughout the day is important for estimating absorption.

The results of absorption estimation are shown on Figure 20 and Table 15. We can see that the precision of class

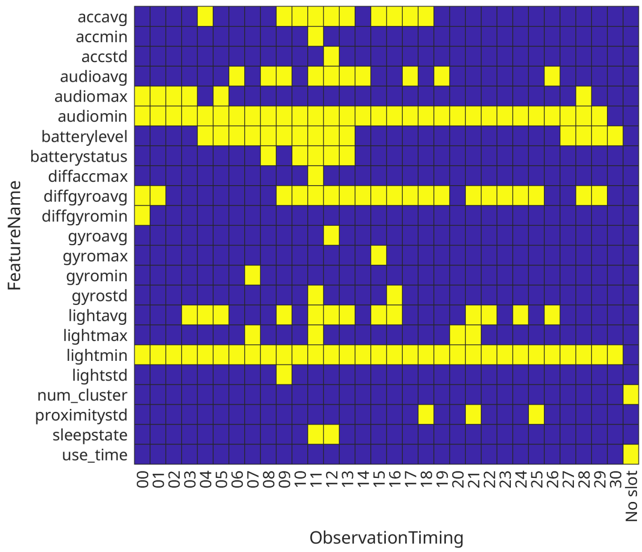


FIGURE 17. Important features of 3-class dedication estimation (Yellow = important, Dark blue = not important). Features without an observation timing are shown in the 'no slot' category.

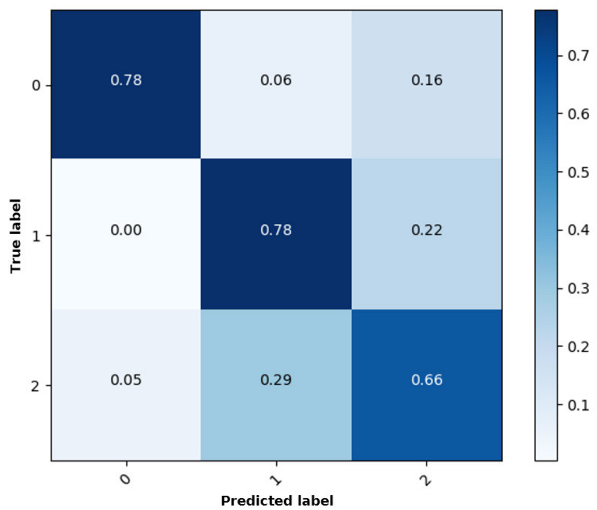


FIGURE 18. Confusion matrix of 3-class dedication estimation (0 = lowest dedication, 2 = highest dedication).

5 is low. This class had insufficient support due to a limited number of training samples, which may have negatively impacted the results. Class 2 also has low precision, despite the high number of samples. Specifically, for true labels in the range of 2-4, the model tends to predict the label 2 incorrectly. However, the model can estimate the extreme cases (0 and 6) very well.

We also evaluated our method with a simplified, 3-class estimation task. The important features are shown on Figure 21, and they are similar to the ones in 7-class estimation, except that the gyroscope features appear to have a stronger influence.

The results are shown on Figure 22 and Table 16. The confusion matrix shows that our model can generally

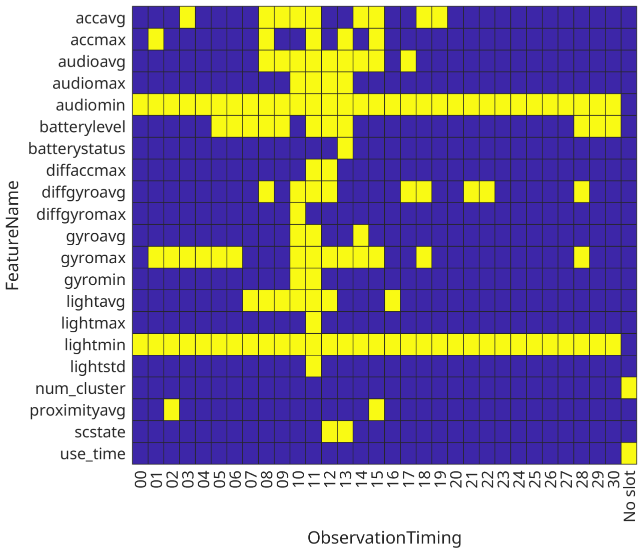


FIGURE 19. Important features of 7-class absorption estimation (Yellow = important, Dark blue = not important). Features without an observation timing are shown in the 'no slot' category.

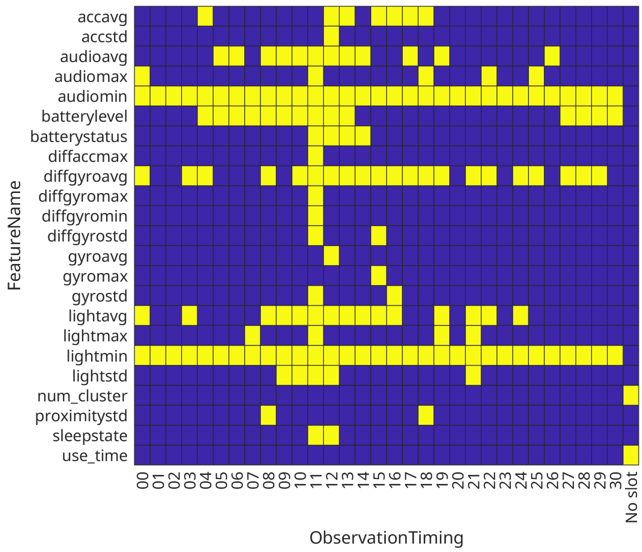


FIGURE 21. Important features of 3-class absorption estimation (Yellow = important, Dark blue = not important). Features without an observation timing are shown in the 'no slot' category.

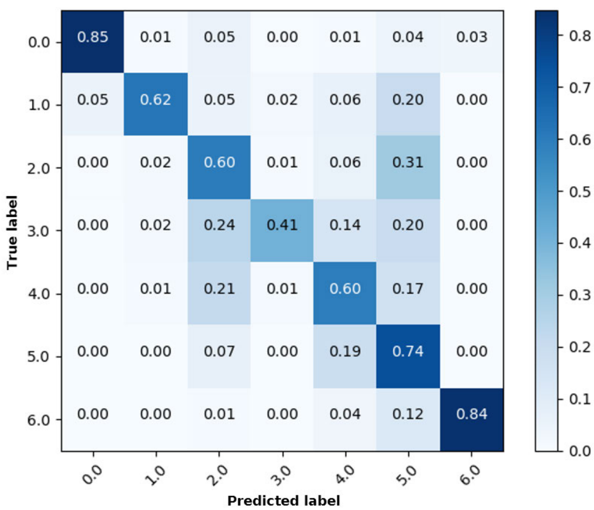


FIGURE 20. Confusion matrix of 7-class absorption estimation (0 = lowest absorption, 6 = highest absorption).

TABLE 15. Evaluation metrics of 7-class absorption estimation.

	precision	recall	f1-score	support
0	0.9692	0.8475	0.9043	223
1	0.7531	0.6162	0.6778	99
2	0.3523	0.5991	0.4437	227
3	0.9733	0.4073	0.5742	717
4	0.5186	0.5953	0.5543	257
5	0.1826	0.7444	0.2932	90
6	0.9307	0.8393	0.8826	112
accuracy			0.5751	1725
macro avg	0.6686	0.6642	0.6186	1725
weighted avg	0.7667	0.5751	0.6081	1725

recognize cases of low absorption well, but has difficulties in distinguishing between neutral and high absorption cases.

As in the 7-class estimation, this is presumably caused by the limited number of training samples for class 5 (in the questionnaire response, on a scale of 7).

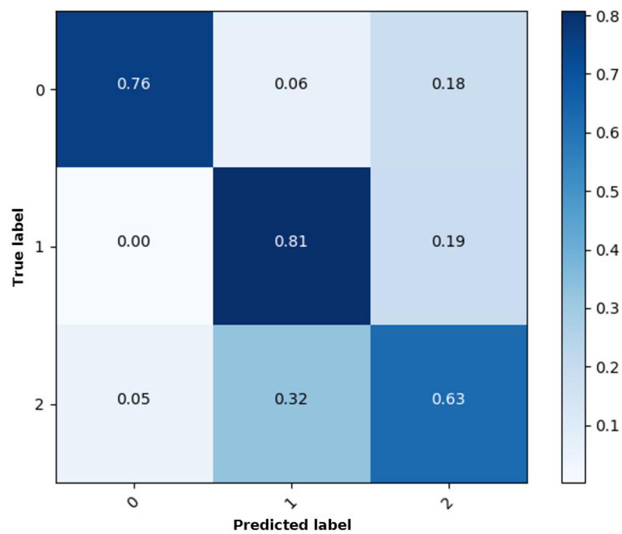


FIGURE 22. Confusion matrix of 3-class absorption estimation (0 = lowest absorption, 2 = highest absorption).

C. DISCUSSION AND LIMITATIONS

In this section, we discuss key limitations of our method and system.

1) TECHNICAL AND SCALING LIMITATIONS

System reliability depends heavily on consistent sensor data collection, which can be compromised by battery drainage or periods of missing/incorrect data when students forget or

TABLE 16. Evaluation metrics of 3-class absorption estimation.

	precision	recall	f1-score	support
0 (Negative)	0.8652	0.7578	0.8079	322
1 (Neutral)	0.7979	0.8085	0.8031	1201
2 (Positive)	0.6047	0.6254	0.6148	202
accuracy			0.7434	1725
macro avg	0.7559	0.7305	0.7420	1725
weighted avg	0.7467	0.7434	0.7444	1725

choose not to carry their phones. As deployments scale to larger student populations, the backend systems will require further optimization to handle increased database loads.

2) PSYCHOLOGICAL AND PRIVACY IMPLICATIONS

The psychological and behavioral implications of continuous monitoring present additional challenges. Students' awareness of being monitored could inadvertently affect their natural behavior patterns - some might experience increased anxiety about being tracked or being identified as "at-risk", while others might attempt to manipulate the system by artificially adjusting their behaviors. Privacy concerns may lead students to restrict app permissions or disable sensing functions, reflecting their psychological discomfort with monitoring. This combination of self-consciousness and privacy-protective behavior could interfere with the very patterns we aim to detect, and potentially worsen social withdrawal tendencies in students already struggling with attendance issues.

3) INDICATOR COVERAGE

Our study did not cover all potential predictors of school attendance issues, and may thus overlook other relevant factors. This could result in false negatives for students affected by unaccounted-for factors. Including additional factors like anxiety, depression, or internet addiction could help reduce these false negatives.

4) GENERALIZABILITY TO OTHER POPULATIONS

Our study was conducted with a limited sample of students sample, offering indicators applicable to those who own and use a smartphone, but generalizability to other populations may require retraining or fine-tuning due to differences in smartphone usage patterns, culture, and resource availability. In low- and middle-income countries with scarce resources like limited Wi-Fi networks and lack of high-end smartphones, deploying our application on a large scale may be challenging. We can adapt by retraining estimation models using available features and sensors, but this may negatively impact accuracy. For instance, if Wi-Fi networks aren't available, the system may use GPS data for location-based features at the expense of privacy, or we can train models without location data. When working with low-cost smartphones with limited sensors, we can train models using the available sensor set. However, our method requires students to have a daily-use smartphone. Our system can be utilized for various age groups, including high school students and even younger students in some cases. With 2022 statistics showing a 45% smartphone penetration rate among 6-12-year-olds in Japan, and over 86% for 13-19-year-olds [34], these users often spend hours daily on their phones [35]. Similarly, our method applies to working adults, as sleep indicators remain relevant, while work engagement replaces study engagement in the prediction model.

5) COMPARISON WITH EXISTING APPROACHES FROM SECTION II

Our approach advances beyond existing work in several key aspects. Regarding *sleep estimation*, existing approaches have shown that smartphone sensors can effectively estimate sleep patterns and quality. However, university students face multiple challenges beyond sleep, necessitating a comprehensive approach that incorporates diverse indicators of well-being. Our work addresses this by incorporating sleep patterns as one component within a broader framework for detecting early warning signs of academic disengagement. In terms of *psychological state estimation*, our work can continuously monitor psychological states and engagement through passive smartphone sensing alone, requiring neither additional devices nor active user participation. This is an advantage over existing methods that rely on wearable devices or active user input, which may not be practical for students with low engagement or social withdrawal tendencies.

VI. CONCLUSION AND FUTURE WORK

In this study, we proposed a screening method and built a student support system that aims to detect subtle signs of school attendance problems in university students. In cooperation with psychiatrists, we defined sleep problems and decreased student engagement as relevant indicators. Unlike existing work, we estimated these indicators using only off-the-shelf smartphones, which are widely used among university students, without requiring additional devices such as wearables.

Our evaluation indicates that, using smartphone sensor data, our Balanced Random Forest model can perform binary sleep state classification (F1-score: 0.808), can estimate subjective sleep quality (3 classes) (F1-score: 0.703), and can estimate 3 standard indicators of study engagement (F1-score: approx. 0.745).

It is important to emphasize that our proposed screening method is not intended to replace traditional face-to-face medical examinations, but rather to complement them by selectively flagging at-risk students and connecting them with medical experts as needed. Ultimately, we believe that our system can have a positive impact on the timeliness and effectiveness of detecting and treating mental health issues such as the initial stages of social withdrawal.

Our next phase of development includes a cross-platform mobile system for Osaka University's 2025 incoming class that will analyze smartphone-generated behavioral data to deliver wellness insights and facilitate optional connections with mental health practitioners.

ACKNOWLEDGMENT

(Viktor Erdélyi and Teruhiro Mizumoto contributed equally to the work.)

REFERENCES

[1] N. Breslau, T. Roth, L. Rosenthal, and P. Andreski, "Sleep disturbance and psychiatric disorders: A longitudinal epidemiological study of young adults," *Biol. Psychiatry*, vol. 39, no. 6, pp. 411–418, Mar. 1996.

[2] C. Uchida, "Apathetic and withdrawing students in Japanese universities-with regard to hikikomori and Student apathy," *J. Med. Dental Sci.*, vol. 57, no. 1, pp. 95–108, Mar. 2010.

[3] C.-Y. Hsu, A. Ahuja, S. Yue, R. Hristov, Z. Kabelac, and D. Katabi, "Zero-effort in-home sleep and insomnia monitoring using radio signals," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–18, Sep. 2017.

[4] Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A. T. Campbell, "Unobtrusive sleep monitoring using smartphones," in *Proc. 7th Int. Conf. Pervasive Comput. Technol. Healthcare Workshops*, May 2013, pp. 145–152.

[5] D. Nagata, Y. Arakawa, T. Kubo, and K. Yasumoto, "Effective napping support system by hypnagogic time estimation based on heart rate sensor," in *Proc. 6th Augmented Human Int. Conf.*, Mar. 2015, pp. 201–202.

[6] M. Hassib, S. Schneegass, P. Eiglsperger, N. Henze, A. Schmidt, and F. Alt, "EngageMeter: A system for implicit audience engagement sensing using electroencephalography," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2017, pp. 5114–5119.

[7] Y. Wang, S. Lu, and D. Harter, "Multi-sensor eye-tracking systems and tools for capturing Student attention and understanding engagement in learning: A review," *IEEE Sensors J.*, vol. 21, no. 20, pp. 22402–22413, Oct. 2021.

[8] E. Di Lascio, S. Gashi, and S. Santini, "Unobtrusive assessment of Students' emotional engagement during lectures using electrodermal activity sensors," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–21, Sep. 2018.

[9] N. Jaques, S. Taylor, A. Azaria, A. Ghandeharioun, A. Sano, and R. Picard, "Predicting students' happiness from physiology, phone, mobility, and behavioral data," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 222–228.

[10] S. Gashi, E. Di Lascio, and S. Santini, "Using Students' physiological synchrony to quantify the classroom emotional climate," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, Oct. 2018, pp. 698–701.

[11] X. Xu, P. Chikersal, A. Doryab, D. K. Villalba, J. M. Dutcher, M. J. Tumminia, T. Althoff, S. Cohen, K. G. Creswell, J. D. Creswell, J. Mankoff, and A. K. Dey, "Leveraging routine behavior and contextually-filtered features for depression detection among college students," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–33, Sep. 2019.

[12] R. Wang, W. Wang, A. daSilva, J. F. Huckins, W. M. Kelley, T. F. Heatherton, and A. T. Campbell, "Tracking depression dynamics in college students using mobile phone and wearable sensing," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–26, Mar. 2018.

[13] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, "StressSense: Detecting stress in unconstrained acoustic environments using smartphones," in *Proc. ACM Conf. Ubiquitous Comput.*, Sep. 2012, pp. 351–360.

[14] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 671–676.

[15] A. Sano, "Measuring college students' sleep, stress, mental health and wellbeing with wearable sensors and mobile phones," Ph.D. dissertation, Program Media Arts Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2016.

[16] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2014, pp. 3–14.

[17] M. Borazio and K. V. Laerhoven, "Predicting sleeping behaviors in long-term studies with wrist-worn sensor data," in *Proc. Int. Joint Conf. Ambient Intell.*, Jan. 2011, pp. 151–156.

[18] M. Littner, C. A. Kushida, W. M. Anderson, D. Bailey, R. B. Berry, D. G. Davila, M. Hirshkowitz, S. Kapen, M. Kramer, D. Loube, M. Wise, and S. F. Johnson, "Practice parameters for the role of actigraphy in the study of sleep and circadian rhythms: An update for 2002," *Sleep*, vol. 26, no. 3, pp. 337–341, May 2003.

[19] F. Wahl and O. Amft, "Data and expert models for sleep timing and chronotype estimation from smartphone context data and simulations," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–28, Sep. 2018.

[20] T. Hao, G. Xing, and G. Zhou, "ISleep: Unobtrusive sleep quality monitoring using smartphones," in *Proc. 11th ACM Conf. Embedded Networked Sensor Syst.*, Nov. 2013, pp. 1–14.

[21] *MIMOSYS (Mind Monitoring System)*. Accessed: Nov. 1, 2022. [Online]. Available: https://medical-pst.com/en/products/mimosys

[22] S. Huynh, S. Kim, J. Ko, R. K. Balan, and Y. Lee, "Engagemon: Multi-modal engagement sensing for mobile games," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, pp. 131–1327, Mar. 2018.

[23] T. Ito, Y. Fukazawa, T. Okimura, Y. Yamashita, T. Maeda, and J. Ota, "Proposal of the prediction model of anxiety using sensor logs and application history," Inf. Process. Soc. Japan, Tokyo, Japan, IPSJ SIG Tech. Rep., vol. 2017-MBL-82, no. 55, Mar. 2017, pp. 1–8.

[24] M. Deng, X. Wang, and C. C. Menassa, "Measurement and prediction of work engagement under different indoor lighting conditions using physiological sensing," *Building Environ.*, vol. 203, Oct. 2021, Art. no. 108098.

[25] G. Urh and V. Pejović, "TaskyApp: Inferring task engagement via smartphone sensing," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Computing: Adjunct*, Sep. 2016, pp. 1548–1553.

[26] A. Turner. *How Many Smartphones Are In The World?*. Accessed: Nov. 7, 2022. [Online]. Available: https://www.bankmycell.com/blog/how-many-phones-are-in-the-world

[27] J. Chan. *Smartwatch Popularity Continues As We Enter the Festive Season*. Accessed: Nov. 7, 2022. [Online]. Available: https://www.kantar.com/inspiration/technology/smartwatch-popularity-continues-as-we-enter-the-festive-season

[28] (2019). *Results of Communication Usage Trend Survey In 2019*. Accessed: Nov. 1, 2022. [Online]. Available: https://www.soumu.go.jp/johotsusintokei/english/

[29] P. Simor, K. N. Krietsch, F. Köteles, and C. S. McCrae, "Day-to-day variation of subjective sleep quality and emotional states among healthy university students—A 1-Week prospective study," *Int. J. Behav. Med.*, vol. 22, no. 5, pp. 625–634, Oct. 2015.

[30] B. Ahammed, N. Jahan, A. Seddeque, M. T. Hossain, T.-E.-A. Shovo, B. Khan, M. A. Mamun, and M. N. Islam, "Exploring the association between mental health and subjective sleep quality during the COVID-19 pandemic among Bangladeshi university students," *Heliyon*, vol. 7, no. 5, May 2021, Art. no. e07082.

[31] A. J. Scott, T. L. Webb, M. Martyn-St James, G. Rowse, and S. Weich, "Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials," *Sleep Med. Rev.*, vol. 60, Dec. 2021, Art. no. 101556.

[32] W. B. Schaufeli, A. Shimazu, J. Hakanen, M. Salanova, and H. De Witte, "An ultra-short measure for work engagement: The UWES-3 validation across five countries," *Eur. J. Psychol. Assessment*, vol. 35, no. 4, pp. 577–591, Jul. 2019.

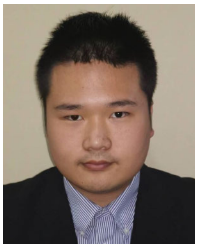
[33] T. Vincenty, "Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations," *Surv. Rev.*, vol. 23, no. 176, pp. 88–93, Apr. 1975.

[34] (2022). *Share of People Who Own Smartphones in Japan in 2022, By Age Group*. Accessed: Aug. 24, 2023. [Online]. Available: https://www.statista.com/statistics/687678/japan-smartphone-penetration-rate-by-age/

[35] *Japan High School Girls Spend 7 Hours a Day on Phones: Survey*. Accessed: Aug. 24, 2023. [Online]. Available: https://www.ctvnews.ca/lifestyle/japan-high-school-girls-spend-7-hours-a-day-on-phones-survey-1.2229021

VIKTOR ERDÉLYI received the B.Sc. and M.Sc. degrees in informatics from Eötvös Loránd University, Hungary, in 2009 and 2011, and the Dr.-Ing. degree in informatics from the Max Planck Institute for Software Systems (MPI-SWS), Germany, in 2019. Since 2019, he has been a specially appointed Assistant Professor with the Graduate School of Information Science and Technology, Osaka University, Japan. He has been working on context-aware systems. His current research interests include mobile and the IoT sensing, wireless (microwave) sensing, object recognition, and localization.

TERUHIRO MIZUMOTO (Member, IEEE) received the B.E. degree from Kindai University, Japan, in 2009, and M.E. and Ph.D. degrees from Nara Institute of Science and Technology, Japan, in 2011 and 2014, respectively. From 2016 to 2018, he was a Postdoctoral Fellow and an Assistant Professor with Nara Institute of Science and Technology. In 2019, he joined the Graduate School of Information Science, Osaka University, as a specially appointed Assistant Professor (full-time). In 2023, he joined Chiba Institute of Technology, while maintaining their position as an Invited Associate Professor at Osaka University. His current research interests include health monitoring, the Internet of Things, and activity recognition. He is a member of IPSJ.



YUICHIRO KITAI received the B.E. and M.E. degrees from Osaka University, Japan, in 2019 and 2021, respectively. He is currently with Nippon Telegraph and Telephone West Corporation.

DAIKI ISHIMARU received the Ph.D. degree from Osaka Prefecture University, Osaka, Japan, in 2020. He is currently an Occupational Therapist with the Department of Psychiatry, Osaka University Hospital. His main research interest includes support for mental health problems.

HIROYOSHI ADACHI received the Ph.D. degree from Osaka University, Osaka, Japan, in 2004. He is currently a Professor with the Health and Counseling Center, Osaka University. His main research interest includes early detection of mental health problems through sleep disturbances.

TERUO HIGASHINO (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in information and computer sciences from Osaka University, Japan, in 1979, 1981, and 1984, respectively. In 1984, he joined as a Faculty Member at Osaka University. Until March 2021, he was a Professor with the Graduate School of Information Science and Technology, Osaka University. In April 2021, he moved to the Faculty of Engineering, Kyoto Tachibana University, as a Professor, where he is currently a Vice President. His research interests include mobile/ubiquitous computing systems, context-aware services, ITS, advanced sensing systems for healthcare, and CPS for disaster mitigation. He is a fellow of the Information Processing Society of Japan. He has served as an Associate Editor for IEEE Transactions on Mobile Computing and *ACM Transactions on Computing for Healthcare*.

MANABU IKEDA received the B.S., M.D., and Ph.D. degrees. He was the Chairperson and a Professor with the Department of Neuropsychiatry, Faculty of Life Sciences, Kumamoto University, from 2007 to 2016. He has been the Chairperson and a Professor with the Department of Psychiatry, Osaka University, since 2016. He has also been engaged with anthropology, neuropsychology, neuropathology, and old age psychiatry research with Tokyo University, Osaka University, Psychiatric Research Institute of Tokyo, Hyogo Institute for Aging Brain and Cognitive Disorders, University of Cambridge, and Ehime University School of Medicine. With his background in neuropsychiatry, neuropsychology, and old age psychiatry, his research covers a wide area of clinical neuropsychiatry.