Personalized Physician-Assisted Sleep Advice for Shift Workers: Machine Learning Approach

Yufei Shen, Alicia Choto Olivier, Han Yu, Asami Ito Matsui, Ryota Sakamoto, Motomu Shimaoka, Akane Sano

Submitted to: Journal of Medical Internet Research
on: August 02, 2024

Table of Contents

Personalized Physician-Assisted Sleep Advice for Shift Workers: Machine Learning Approach

Yufei Shen[1]; Alicia Choto Olivier[1]; Han Yu[1]; Asami Ito Matsui[2]; Ryota Sakamoto[2]; Motomu Shimaoka[2]; Akane Sano[1]

[1]Rice University Houston US
[2]Mie University Tsu JP

Corresponding Author:
Akane Sano
Rice University
6100 Main St
Houston
US

Abstract

Background: In the modern economy, shift work is prevalent in numerous occupations. However, shift work often conflicts with the workers' circadian rhythm and can result in shift work sleep disorder (SWSD). Proper management of SWSD emphasizes comprehensive and patient-specific strategies and some of these strategies are analogous to the cognitive behavioral treatment of insomnia (CBTI).

Objective: In this paper, we aim to develop and evaluate machine learning algorithms that predict physicians' sleep advice using wearable and survey data. We developed an online system to conveniently and frequently provide individualized sleep and behavior advice with CBTI elements for shift workers.

Methods: Data were collected for a period of 5 weeks from shift workers in the ICU at two hospitals (N = 61) in Japan. The data were composed of three modalities, (1) Fitbit data, (2) survey data, and (3) sleep advice. We handcrafted physiological and behavioral features from the raw data and identified clusters of participants with similar characteristics using hierarchical clustering. After the first week of enrollment, physicians reviewed Fitbit and survey data to provide sleep advice from a list of 23 messages. We implemented random forest (RF) models to predict the 7 most frequent messages given by the physicians. We tested our predictions under participant dependent and independent settings and analyzed the most important features for prediction.

Results: We found that the clusters were distinguished by work shifts and behavioral patterns. For some clusters, having a work shift on a given day contributed to low wellbeing scores on that day. Another cluster had days with low sleep duration and the lowest sleep quality when there was a day shift on the day before and a midnight shift on the current day. Our advice prediction models achieved higher F1 scores in 27 of 28 t-tests conducted, and the performance differences were statistically significant with $P < .001$ for 24 tests and $P < .05$ for 3 tests compared to the baseline. The analysis of the feature importance of our models showed that the most important features matched the message sent to participants. For instance, for message 7 (darken the bedroom when you go to bed), the models primarily examined the average brightness of the sleep environment to make predictions.

Conclusions: Although our current system requires physician input, an accurate machine learning algorithm would be promising for automating without hurting the trustworthiness of selected recommendations. The algorithm is limited to the 7 most popular ones among 23 choices due to rare occurrences of the remaining options. Therefore, further studies are necessary to gather enough data to enable predictions for less frequent advice labels. Clinical Trial: UMIN Clinical Trials Registry UMIN000036122 (phase 1), UMIN000040547 (phase 2); https://tinyurl.com/dkfmmmje, https://upload.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000046284.

(JMIR Preprints 02/08/2024:65000)
DOI: https://doi.org/10.2196/preprints.65000

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Personalized Physician-Assisted Sleep Advice for Shift Workers: Machine Learning Approach

Yufei Shen [1], Alicia Choto Olivier [1], Han Yu [1], Asami Ito [2,3], Ryota Sakamoto[4], Motomu Shimaoka[5], Akane Sano [1]

1) Department of Electrical & Computer Engineering, Rice University
2) Emergency and Critical Care Center, Mie University Hospital
3) Department of Emergency & Disaster Medicine, Mie University Graduate School of Medicine
4) Department of Medical Informatics, Mie University Hospital, Tsu, Japan
5) Department of Molecular Pathology & Cell Adhesion Biology, Mie University Graduate School of Medicine, Tsu, Japan

Abstract

Background:

In the modern economy, shift work is prevalent in numerous occupations. However, shift work often conflicts with the workers' circadian rhythm and can result in shift work sleep disorder (SWSD). Proper management of SWSD emphasizes comprehensive and patient-specific strategies and some of these strategies are analogous to the cognitive behavioral treatment of insomnia (CBTI).

Objective:

In this paper, we aim to develop and evaluate machine learning algorithms that predict physicians' sleep advice using wearable and survey data. We developed an online system to conveniently and frequently provide individualized sleep and behavior advice with CBTI elements for shift workers.

Methods:

Data were collected for a period of 5 weeks from shift workers in the ICU at two hospitals (N = 61) in Japan. The data were composed of three modalities, (1) Fitbit data, (2) survey data, and (3) sleep advice. We handcrafted physiological and behavioral features from the raw data and identified clusters of participants with similar characteristics using hierarchical clustering. After the first week of enrollment, physicians reviewed Fitbit and survey data to provide sleep advice from a list of 23 messages. We implemented random forest (RF) models to predict the 7 most frequent messages given by the physicians. We tested our predictions under participant dependent and independent settings and analyzed the most important features for prediction.

Results:

We found that the clusters were distinguished by work shifts and behavioral patterns. For some clusters, having a work shift on a given day contributed to low wellbeing scores on that day. Another cluster had days with low sleep duration and the lowest sleep quality when there was a day shift on the day before and a midnight shift on the current day. Our advice prediction models achieved higher F1 scores in 27 of 28 t-tests conducted, and the performance differences were statistically significant with $P < .001$ for 24 tests and $P < .05$ for 3 tests compared to the baseline. The analysis of the feature importance of our models showed that the most important features matched the message sent to participants. For instance, for message 7 (darken the bedroom when you go to bed), the models primarily examined the average brightness of the sleep environment to make predictions.

Conclusions:

Although our current system requires physician input, an accurate machine learning algorithm would be promising for automating without hurting the trustworthiness of selected recommendations. The algorithm is limited to the 7 most popular ones among 23 choices due to rare occurrences of the remaining options. Therefore, further studies are necessary to gather enough data to enable predictions for less frequent advice labels.

Trial Registration: UMIN Clinical Trials Registry UMIN000036122 (phase 1), UMIN000040547 (phase 2); https://tinyurl.com/dkfmmmje, https://upload.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000046284.

Keywords: CBT; cognitive behavioral therapy; health care workers; machine learning; medical safety; online intervention; shift work; shift work sleep disorders; shift workers; sleep; sleep disorder; wearable sensors; well-being.

Introduction

Background

In the modern economy, shift work is prevalent in numerous occupations. Data from 2010 estimated that half of the workers in food preparation and serving, more than 45% of workers in protective service, and more than 35% of healthcare practitioners had alternative shifts apart from regular day shifts in the United States [1]. However, shift work often conflicts with the daily rhythm of sleep and wakefulness, which increases the risk of shift work sleep disorder (SWSD), a circadian rhythm disorder characterized by shift work-related sleepiness and insomnia [2]. Multiple studies have shown the substantial prevalence of SWSD among shift workers. A survey of nurses in 3 federal hospitals in Ethiopia found that 25.6% (102/399) of participants had SWSD [3]. Another study of nurses in Norway reported prevalence rates of 44.2% (580/1313) and 23.6% (146/619) for SWSD indicative symptoms among night workers and day workers respectively [4]. A random population sample of 1163 participants in Australia revealed that 10.1% of day workers (91/898) and 32.1% of night workers (85/265) fulfilled the study's criteria for SWSD, and 1.3% (12/898) of day workers and 9.1% (24/265) of night workers were further classified under severe SWSD based on the extent of negative SWSD impacts on their life [5]. Shift workers' sleep problems are also associated with risks of mental health issues [6–8], work errors [8], burnout [6], and turnover intentions [9].

Proper management of SWSD emphasizes comprehensive and patient-specific strategies such as circadian adaptation, maintaining good sleep hygiene, strategic napping, clockwise shift rotation, and wakefulness promotions for night shifts [10]. Some of these strategies are analogous to the cognitive behavioral treatment of insomnia (CBTI), a psychological therapy recommended by the American College of Physicians as initial treatment for chronic insomnia [11] which consists of multiple approaches to improve sleep including education about normal sleep, sleep hygiene practice, and sleep restriction to consolidate high quality sleep, and relaxation [12].

Related Work

Researchers have experimented with CBTI on shift workers with sleep problems and high risks of SWSD. A study by J¨arnefelt et al on 19 shift workers with insomnia that more than half of whom exhibited SWSD characteristics showed that group-based CBTI treatment improved their insomnia severity and sleep over a 6-month post-intervention period [13], and a follow-up study further demonstrated long-lasting benefits of CBTI for up to 24 months after interventions [14]. J¨arnefelt et al also conducted a randomized controlled trial of group-based CBTI and self-help CBTI against sleep hygiene education control on 83 shift workers with insomnia more recently. They found that the group-based CBTI improved participants' mental health significantly although both CBTI interventions enhanced their sleep and insomnia symptoms without significant differences from the control and all interventions were more beneficial for participants without SWSD characteristics [15]. Peter et al employed an online CBTI for shift workers with sleep problems (N=21) and found it to be as effective as face-to-face SWSD outpatient CBTI treatment (N=12) for improving sleep efficiency and insomnia symptoms [16]. Their research group also proposed a randomized controlled trial in 2021 to further assess their online CBTI intervention for shift workers against a waitlist control and face-to-face CBTI [17]. Although existing research on CBTI for shift workers showed promising improvement in their sleep, group-based and face-to-face CBTI required participants to attend several weekly sessions for up to 120 minutes led by specialists trained for more than 10 hours, and the long time commitment likely caused many participants to miss some sessions which might weaken intervention effects [13,15,16]. Self-help online CBTI was relatively more accessible to shift workers, but time commitment for weekly interventions was still necessary and likely contributed to the considerable dropout rates, and the interventions either required human input for professional feedback or only resembled an online learning module without any personalization [15,16]. Therefore, a system that automatically provides personalized CBTI interventions and sleep recommendations for shift workers with less time commitment could benefit them substantially and yield low dropout rates.

In recent years, several online and mobile CBTI systems have been developed to deliver automated and individualized CBTI interventions. For example, SHUTi [18] and Sleepio [19] provided automated online CBTI modules equipped with interactive web design elements and user-tailored recommendations, and multiple randomized controlled trials have demonstrated their effectiveness against insomnia and related mental health problems among various demographic groups [18,20–23]. Beun et al developed Sleepcare, a mobile app equipped with an automated digital coach interacting with users to provide CBTI interventions, adapt interventions to user characteristics, and promote adherence [24,25]. The efficacy has been manifested in a randomized controlled trial of 151 participants by Horsch et al [26]. CBT-I coach was another phone app developed as a supplemental tool for face-to-face CBTI in which users could receive sleep education, customize sleep hygiene advice and relaxation activities, and practice personalized sleep restrictions [27]. In a pilot study, the app did not impair CBTI effectiveness and potentially improved patient adherence [28]. A subsequent variant of the app called Insomnia Coach was applied as a self-help CBTI on 25 veterans with insomnia and showed significant treatment effects compared to the wait list control [29].

Besides online and mobile CBTI, multiple automated sleep recommender systems (RS) have also been developed to provide personalized sleep suggestions. Sleepcoacher collected data from smartphone sensors and user self-reports to compute sleep variables and included recommendation templates reviewed by clinicians for each variable combination [30]. To deliver a recommendation, the system chose the template of the most correlated combination and further personalized it according to the user's sleep status, and the system also evaluated the effectiveness of the recommendation. Two studies on 43 undergraduate students showed the benefits of the system on sleep which further increased with higher adherence rates. Daskalova et al explored cohort-based sleep recommendations where their system used demographic, exercise, and sleep data collected from 1 million users of a wearable device to form a user group with similar profiles for each participant, and chose a recommendation that have shown the greatest improvement among the group in some aspects of sleep behaviors which the participant performed poorly compared to these similar users [31]. Compared to general sleep suggestions, cohort-based recommendations improved sleep duration more, and many participants were motivated to follow these recommendations. Pandey et al utilized event mining to discover causal relationships between a user's lifestyle and sleep, and provided recommendations to match their current behaviors and environment with these optimal relationships for better sleep [32]. On the other hand, some systems did not directly provide advice to users but instead helped them to self-adjust their sleep habits. For example, Lullaby utilized multiple sensors to track users' sleep and environment and facilitated them to discover environmental factors with negative impacts on their sleep [33]. ShutEye served as a peripheral display on mobile phones to remind users of recommended and inadvisable time windows for various sleep-related activities [34]. SleepBandits enabled users to perform numerous self-experiments to adjust their behaviors and used data from mobile phone sensors to evaluate experiment effects on their sleep [35].

Although existing automated and individualized CBTI and sleep RS have demonstrated their great potential in improving sleep, they were not modified to accommodate shift workers' irregular schedules of shift and wakefulness [13] and thus did not include strategies specifically beneficial to shift workers. Moreover, these CBTI interventions were mostly delivered through weekly learning modules and tasks requiring active involvement and regular time commitment, and their personalization options were limited and mainly focused on sleep restrictions. Unlike CBTI, sleep RS directly sent behavior change recommendations to users, but their advice tailoring processes often lacked clinical support and raised doubts about their credibility among users [31].

Objective

In this paper, we aim to develop and evaluate a machine learning algorithm that predicts physicians' sleep advice using wearable and survey data. We also examined the model's potential in assisting and even replacing clinical reviews in the system.

We characterized shift workers' behaviors using clustering analyses, estimated physicians' advice selection strategies by interpreting the developed advice prediction models, and assessed the effectiveness of provided sleep recommendations to gain more insights into the system.

Methods

Data Collection

The data were collected from shift workers in ICU at Mie University Hospital (N = 47) and Suzuka Chuo General Hospital (N=14) in Japan [36]. The protocol was approved by Mie University. Detailed inclusion and exclusion criteria were described by Ito-Masui et al [36]. Each participant enrolled in a 5-week trial with the first week as a baseline without intervention (pre-intervention) and the following 4 weeks with interventions. During the 5 weeks, participants filled out daily surveys and wore a Fitbit to collect physiological and behavioral data. Physicians reviewed the data and provided sleep advice to participants as interventions 3-4 times a week. In total, 2123 days of multimodal data were collected from these participants for the entire study.

Data Modalities

The collected data included three main modalities - Fitbit data, survey data, and sleep advice. We followed Yu et al's methods [37] to extract some features from the same Fitbit and survey data, and engineered extra features to accommodate our objectives.

Fitbit Data

Raw data from Fitbit included minute-by-minute steps and heart rate as well as start time, end time, duration, and efficiency of each sleep period. From raw heart rate data, we computed its average, standard deviation, and sample entropy. Sample entropy has been used for cardiovascular time series and a low entropy value suggests greater self-similarity of the series [38]. We also computed features from step counts per minute to account for the variability of participants' daily activities. Specifically, we computed and stored the duration of each continuous segment with and without steps in minutes. Then we computed information entropy for the stationary and active segments. A higher entropy corresponded to more variability of participants' moving or non-moving behaviors.

Finally, we retrieved information about the main sleep period with the longest duration from raw sleep data.

Survey Data

Participants received a morning survey and an evening survey every day with questions about their daily behaviors and wellbeing. Features were then extracted from their answers.

Morning Survey

Features from morning surveys are categorized into three categories - sleep, wellbeing, and miscellany.

Sleep

In morning surveys, participants indicated whether they slept in the previous 24 hours and provided the numbers of naps between 8am and 8am on the following day. They also described how they woke up (naturally, by alarm, or by other means), and reported the time it took for them to fall asleep, the duration of their phone use prior to sleep, the brightness level of their sleep environment, and sleep quality. Specifically, sleep quality was quantified by answering the following statements using 5-point Likert scales (strongly disagree (1) to strongly agree (5)): (a) I slept soundly, (b) I fell asleep immediately, (c) I was able to recover from fatigue, (d) I didn't wake up in the middle of sleep, (e) I was satisfied with sleep.

Besides the above features retrieved directly from surveys, we also constructed some sleep features using available survey data. In the surveys, participants reported the start time and end time of their main sleep and nap periods. According to the time entries, we constructed a binary sequence with one minute resolution for every day, where a bit of one meant that the participants were sleeping or napping at that minute of the day. If time entries of main sleep periods were missing from the surveys, the entries from Fitbit were used. From the binary sequences, we computed the duration of main sleep and nap periods. We also calculated Sleep Regularity Index (SRI) with sliding windows of 3, 5, and 7 days. SRI represents the probability of any two time points that are 24 hours apart having the same sleep or awake state averaged over a specified time window, and is scaled to range from -100 to 100, where a value of 100 indicates the same sleep schedule across all days and a value of -100 means that an individual has completely flipped sleep schedule between any two consecutive days [39]. Previous studies have shown that SRI is positively correlated with academic performance of college students [39], and lower SRI is associated with increased stress, depression, and risk of cardiovascular diseases in older adults [40].

Wellbeing

Participants reported 6 wellbeing metrics in the morning surveys. Five metrics including alertness, happiness, energy, health, and calmness were reported with a continuous scale from 0 to 100 with 100 as the most positive. Another metric of current sleepiness was reported by a 9-point Likert scale from strongly awake (1) to strongly sleepy (9).

Behavior

Morning surveys also recorded participants' daily behaviors including the amount of caffeine intake and last intake time, amount of alcohol intake, and bath time. From the time entries, we computed durations between their last caffeine intake and sleep, and between bath and sleep.

Evening Survey

Features from evening surveys are categorized into two categories - work and wellbeing.

Work

In evening surveys, participants reported their work schedules for today, yesterday, and the day before yesterday in binary sequences with 30-minute resolution, where a bit of one meant work during that 30-minute interval. From the binary sequences, we computed work hours for the entire day as well as for three different periods of the day (1: 0:00 - 8:00, 2: 8:00 -16:00, and 3: 16:00 - 0:00). The periods corresponded to midnight shift, day shift, and afternoon shift of participants. Moreover, by comparing the binary work sequences and binary sleep sequences constructed from morning surveys, we calculated participants' nap durations during work, minimum durations between the last sleep periods and the start of shifts, and between the end of shifts and following sleep periods. A binary indicator of whether participants did any activities other than work outside their homes was also retrieved from the surveys.

Wellbeing

Similarly, as in morning surveys, participants also answered several questions about their wellbeing in evening surveys. Besides the 6 metrics reported in morning surveys (alertness, happiness, energy, health, calmness, and current sleepiness), they also indicated three additional metrics including sleepiness during the day, stress, and tiredness, which were all reported by 5-point Likert scales (5 for strongly awake, stressful, or tired).

Sleep Advice

Starting from the second week of enrollment, physicians reviewed Fitbit and survey data to provide sleep advice. One physician provided advice once a week to all participants, and two physicians gave advice 2-3 times a week to participants from two hospitals respectively. Over the 4 weeks with interventions, participants received advice around 13 times, and 786 pieces of advice were sent to the 61 participants who completed their 5 weeks of trials. Every time physicians provided sleep advice, they chose 1-5 messages from 23 options listed below (Table 1). To better summarize the options, we divided them into 6 categories according to their descriptions- 1) dietary intake (messages 1, 2, and 3), 2) activity (messages 4, 5, 6, 16, 17, and 18), 3) sleep (messages 7, 8, 9, 10, 20, 21, and 22), 4) shift (message 11), 5) nap (messages 12, 13, 14, 15, and 23), and 6) mentality (messages 19). After participants received the sleep advice, they were able to respond to the advice as "eager to follow" or "difficult to follow".

Table 1. A list of messages provided to the participants

Message Category	Message ID	Message Description
Dietary intake		
	1	Refrain from consuming alcohol before sleep.
	2	Stop consuming caffeine 3 hours before lights-out.
	3	Refrain from eating midnight snacks.
Activity		
	4	Refrain from using a smartphone in the bedroom.
	5	Take a bath a little earlier than usual.
	6	Refrain from exercising 3 hours before lights-out.

	16	Relax. 1) Make a habit of relaxing before sleep, including taking a bath in warm water, light stretching, using aroma, and drinking herbal tea.
	17	Relax. 2) Use the Fitbit breathing program called Relax.
	18	Do moderate exercise regularly.
Sleep		
	7	Darken the bedroom when you go to bed.
	8	When you do not fall asleep within 15 minutes, leave the bedroom and stay out of the bedroom until you get sleepy.
	9	Set your wake-up time according to each work shift.
	10	Is there a possibility that you are trying too hard to fall asleep quickly?
	20	Continue current sleep habits.
	21	Try to make enough time for sleep.
	22	Get up at the same time, whether you are working or on holiday.
Shift		
	11	Management of work shift: Consult with your superior so that shift rotation will be clockwise in general.
Nap		
	12	Take a nap. 1) If possible, take a nap for approximately 90 minutes before the night shift.
	13	Take a nap. 2) If possible, take a nap for 15–20 minutes during rest time while you are working.
	14	Take a nap. 3) If possible, take a nap earlier after a late-night shift to refresh.
	15	Create an environment for taking a nap: Ask your family for cooperation to create a quiet environment during a nap in the daytime or evening at home. Sharing your work shift schedule with your family by placing it where every member of your family can see it might be a good idea. Avoid strong lights for several hours before a nap and darken the room during a nap.
	23	Do not nap for too long.
Mentality		
	19	Be broad-minded and try to approach things positively.

Participant Characteristics Analyses

We investigated the characteristics of participants' daily activities to find participants' subgroups who have similar behaviors and understand the relationship between their behaviors and wellbeing. We used agglomerative hierarchical clustering to find the groupings. The algorithm initiates all samples as individual clusters and merges clusters with Ward linkage to minimize the within-cluster variances at every step [41,42]. After all the samples are merged, a dendrogram is constructed to show the merged path for each sample. By setting thresholds on the height of the dendrogram, clusters and their samples can be identified.

We used features from Fitbit and surveys, except for the participants' wellbeing and sleep quality metrics, to find the clusters. To include as many samples as possible during clustering, we did not use the following features since they were not available for some days: SRI, nap durations during work, the time between caffeine intake and sleep, between bath and sleep, between last sleep and the start of shifts, and between the end of shifts and following sleep (e.g., participants did not work or take baths).

We then overlaid the clusters' features on a 2D t-Distributed Stochastic Neighbor Embedding (t-SNE) plot to observe how clusters interact with each other. t-SNE creates a low-dimensional mapping of high-dimensional data by matching the conditional probabilities in both spaces which are proportional to the similarity of data points [43]. To characterize participants' behaviors and wellbeing, we summarized the clusters' profiles by observing feature averages and examined distributions of the wellbeing metrics for each cluster.

Advice Prediction Models

As mentioned in the section Sleep Advice, physicians reviewed the Fitbit and survey data to provide sleep advice to participants. We developed machine learning models to predict their message selections and evaluated the performance of the advice prediction models.

Prediction Approach

We formulated this advice prediction task as a binary classification label for each message as correlations between any two messages are not significant. Given the imbalanced frequencies of messages, we decided to build models only for the 7 most frequent advice (message ID: 4, 7, 12, 14, 15, 20, and 21) because models for less frequent messages would have too few positive samples. The number of total occurrences for the 7 messages is 786, and the individual occurrences ranged from 64 (8.1%) to 390 (49.6%).

We computed the average and standard deviation of daily fitbit and survey features across the previous 4 days for each date on which physicians provided advice. According to the physicians, they often considered the participants' responses to previous advice, and in occasions they found that it was difficult to choose the messages. Therefore, we incorporated advice responses into our models and enabled models to output class probabilities to indicate certainty about a message selection. We used 12 fitbit features (steps, sleep, heart rate), 52 morning survey features, 50 evening survey features and 23 advice response features. Table S1 summarizes features used by the classifier.

Classifier

We used random forest (RF) classifiers for advice predictions. RF aggregates predictions of individual decision trees constructed with only subsets of all features and is robust to overfitting [44]. Because of its robustness, it is suitable for advice predictions with limited samples and many features. Due to class imbalances of message selections, we balanced the RF by undersampling the majority class as implemented by the imbalanced-learn library in Python [45,46].

Hyper-parameters of the RF included the number of decision trees, the maximum depth of trees, the minimum number of samples required to split a node of the tree, the minimum number of samples required at a leaf, and the proportion of all features used to split a node. To find the optimal set of hyper-parameters, we tuned the classifier on the training set with 5-fold cross-validation and 300 iterations of random search. Random search was used instead of grid search because it can span a larger parameter space and find models equally good as or better than grid search results under the same computation budget [47]. In our case, possible values of the hyper-parameters were 20 to 200 with step size 10 for the number of trees, 2 to 30 with step size 3 for maximum depth, 5 to 70 with step size 5 for the minimum number of samples required to split, 5 to 70 with step size 10 for the minimum number of samples required at a leaf, and 0.05 to 0.5 with step size 0.05 for the proportion of all features.

Prediction Evaluation

We evaluated the models under both user dependent and independent settings. Under the user dependent setting, the data was split chronologically where the training set contained the first 70% of each participant's data and the testing set had the remaining 30%. Under the user independent setting, we split the data according to the order of participant enrollment that models were trained on data from the first 70% of participants enrolled and tested on the 30% of participants who enrolled later in the study. We chose the train-test split methods to ensure that when making predictions for a date of a participant, no data from later days of this participant's enrollment would be used in dependent models and only data from participants enrolled earlier would be used in independent models.

We repeated the tuning and training process 10 times and compared the average performance of the classifier (F1 score, precision, and recall) to baselines: an average of 10 random guessing trials, and a classifier that always predicts the positive class. We also measured feature importance by computing decreases in F1 scores on out-of-bag samples of training data after randomly permuting a specific feature. Permutations of important features would lead to greater decreases in F1 scores, as described by Breiman et al [48]. We permuted each feature 10 times for each running of the classifier and summarized the results from all 100 trials. Since decision trees are well-known for their great interpretability by decomposing complex decisions into simpler ones and visualizing decision rules in tree structures [49], we explored the roles of the features in decision making by plotting and examining individual decision trees of fitted RF.

Results

Participant Characteristics

Tables 3 and 4 summarize the participants' daily features and wellbeing scores respectively. Their average step count was 7539 (SD: 2715), sleep duration was 6.82 hours (SD: 64.96 mins), nap duration was 50. 7 mins (SD: 30.2 mins), sleep regularity was 51.7 (SD: 10.5) and average work hour was 5.57 (SD: 1.22) hours.

We used 77% (1635/2123) of samples with all selected features available for hierarchical clustering. Figure 1 shows the resulting dendrogram. The default distance threshold of 52.24 (purple dashed line) identified 6 clusters color-coded and labeled by numbers 0-5 in the dendrogram. Additionally, a few smaller clusters merge at around 30, so we used a lower threshold of 31 (red dashed line) to divide the 6 clusters into 13 subclusters (a-m) and further investigated the behavior variability within each larger cluster. Figure S1 illustrates the clusters and subclusters overlaid on t-SNE plots obtained from the same data used for clustering. Cluster and subcluster structures were clearly observable in both plots. For example, cluster 0 in the plot of 6 clusters has a left branch and a right branch on either side of cluster 1, which are further divided into subcluster g and c respectively in the plot of 13 subclusters.

Table S2 summarizes cluster and subcluster profiles for different aspects of participants' behaviors and wellbeing. Major differences among the 6 clusters were observed in the wellbeing metrics, while subclusters can also be distinguished by work hours and shift types. The profiles will be discussed in detail later in the Discussion section.

Table 3 Participant’s daily features summary statistics

Feature	Mean	SD
hrmean	74.74	6.7
hrstd	13.39	1.84
hrentropy	0.63	0.09
steps	7538.5	2715.47
duration_entropy_non_step	2.36	0.15
duration_entropy_step	1.69	0.16
nap_count	0.57	0.28
brightness_sleep	19.22	15.12
sleep_duration [mins]	409.72	64.96
nap_duration [mins]	50.65	30.16
sleep_regularity_3 days	51.94	10.45
sleep_regularity_5 days	51.61	10.67
sleep_regularity_7 days	51.7	10.47
deep_sleep (scale: 1-5)	3.17	0.57
immediate_sleep (scale: 1-5)	3.52	0.58
fatigue_recover (scale: 1-5)	2.82	0.58
mid_awake (scale: 1-5)	3.22	0.73
sleep_satisfy (scale: 1-5)	2.83	0.54
alcohol_amount	0.31	0.43
caffeine_amount	1.58	1.28
worktime_today_duration [hr]	5.57	1.22
worktime_today_duration (1: 0:00 - 8:00) [hr]	1.47	0.73
worktime_today_duration (2: 8:00 -16:00) [hr]	2.1	0.81
worktime_today_duration (3: 16:00 - 0:00) [hr]	2	0.77
	(%)	
extrawork_activities	43.08	
sleep_prev_24	100	
wake_natural	39.68	
wake_alarm	37.10	
wake_other	1.64	

Table 4 Participant’s wellbeing summary scores.

Wellbeing	**Mean**	**SD**
sleepiness_morn (scale: 1-9)	5.54	0.87
sleepiness_eve (scale: 1-9)	5.72	0.81
sleepiness_daytime (scale: 1-5)	2.56	0.43
Stress (scale: 1-5)	2.90	0.63
Tiredness (scale: 1-5)	2.51	0.50
alertness_morn (0-100)	40.56	11.68
happiness_morn	52.53	10.55
energy_morn	43.85	12.38
health_morn	51.59	12.12
calmness_morn	55.86	11.70
alertness_eve	40.79	10.65
happiness_eve	54.05	10.35
energy_eve	45.73	12.54
health_eve	52.05	12.41
calmness_eve	56.50	12.54

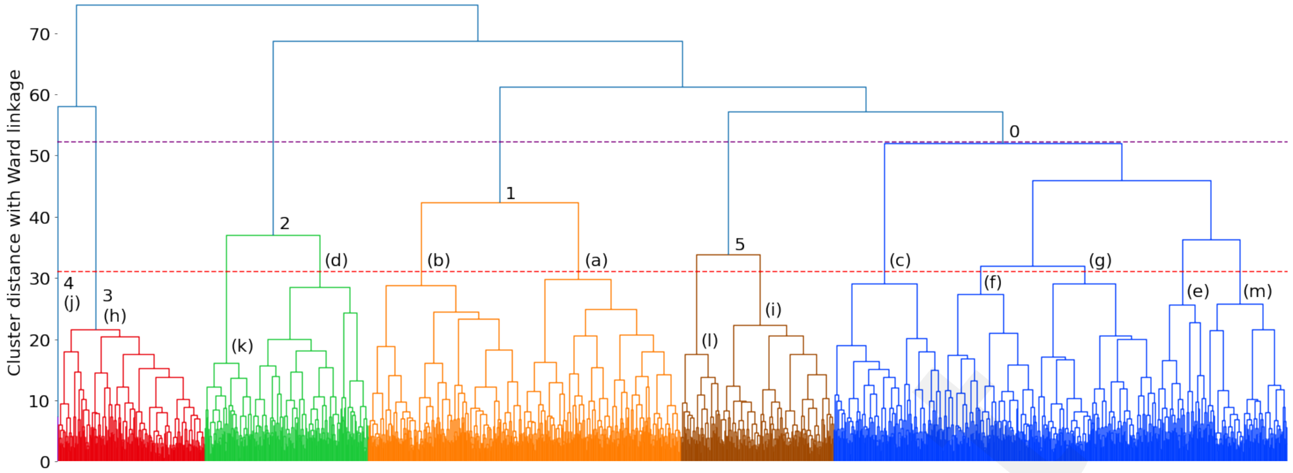

Figure 1: Dendrogram from hierarchical clustering with Ward linkage. The purple dashed line corresponds to a distance threshold of 52.24 (0.7 times maximum cluster distance as implemented by SciPy [1]) and identifies 6 clusters labeled and color-coded by cluster 0 (blue), 1 (orange), 2 (green), 3 (red), 4 (purple), and 5 (brown). The red dashed line represents a distance threshold of 31 and identifies 13 subclusters (a-m). Specifically, cluster 0 has 5 subclusters: c, f, g, e, and m. Cluster 1 has 2 subclusters: a and b. Cluster 2 has 2 subclusters: k and d. Cluster 3 and 4 do not have any subclusters and themselves are labeled as j and h respectively. And cluster 5 has 2 subclusters: l and i.

Advice Statistics

Figure 2 shows the frequency of advice provided as well as the message categories. Since physicians often selected multiple messages for each day, 1332 advice messages were selected in 786 advice pieces. Messages in the sleep and nap categories were selected more frequently than other categories. Messages in the sleep category were selected the greatest number of times (817/1332, 61.34%), and messages in the nap category were selected 289 times (21.70%). Moreover, only a few messages were frequently selected by physicians. For example, message 20 (Continue current sleep habits.) was selected in 390 out of 786 advice pieces (49.6%), and message 21 (Try to make enough time for sleep.) was selected in 221 pieces (28.1%). On the other hand, 16 out of 23 messages were selected in fewer than 50 advice pieces (6.4%), and 7 messages were selected fewer than 10 times (1.3%).

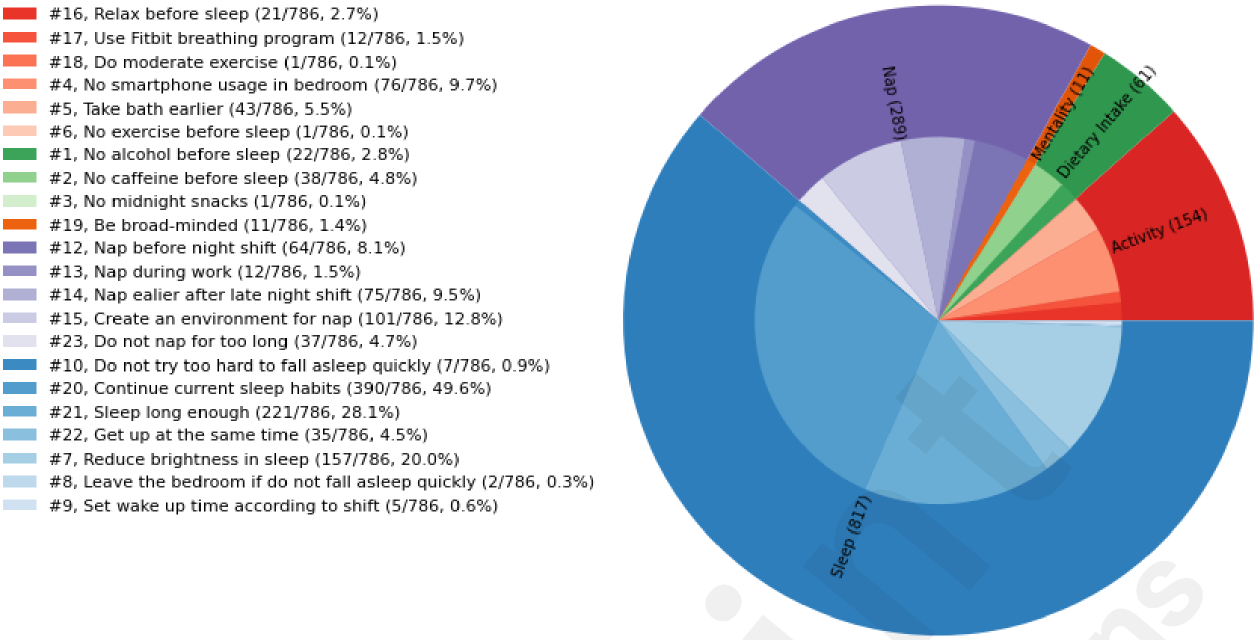

Figure 2: The pie chart shows the frequency of each message. All 23 message options are divided into 6 categories- dietary intake (messages 1, 2, 3), activity (messages 4, 5, 6, 16, 17, 18), sleep (messages 7, 8, 9, 10, 20, 21, 22), shift (message 11), nap (messages 12, 13, 14, 15, 23), and mentality (message 19). Message 11 is excluded from the chart because of zero frequency. All message options were selected 1332 times in 786 advice pieces. The outer ring of the pie chart indicates the proportions of each category with respect to 1332 total occurrences of all message options and the parenthesis after each category name corresponds to the frequency of the category. The inner ring shows the proportion of each message with respect to total occurrences. The legend lists the corresponding color of each message and each parenthesis includes the proportion of all advice pieces with the message selected, which is computed by dividing the frequency of the message by 786 total advice counts.

Figure 3 summarizes the frequency and proportion of responses to advice pieces containing each message. Sometimes participants did not respond to the advice. Although multiple messages might be selected for one advice piece, participants responded to the advice but not to individual messages. Participants showed different preferences for messages. For example, participants expressed difficulty following message 20 (Continue current sleep habits.) only 0.5% of the time (2/390), while they felt difficulty following message 4 (Refrain from using a smartphone in the bedroom.) for 19.7% of the time (15/76), and message 2 (Stop consuming caffeine 3 hours before lights-out) for 26.3% of the time (10/38). The discrepancies might be explained by the message design. Message 20 suggests participants continuing current sleep habits, which participants might feel easier to follow as it would not change their current lifestyles greatly. On the other hand, message 2 and message 4 address caffeine consumption and phone usage, which might be challenging to follow as they were suggested to change some essential aspects of participants' routine.

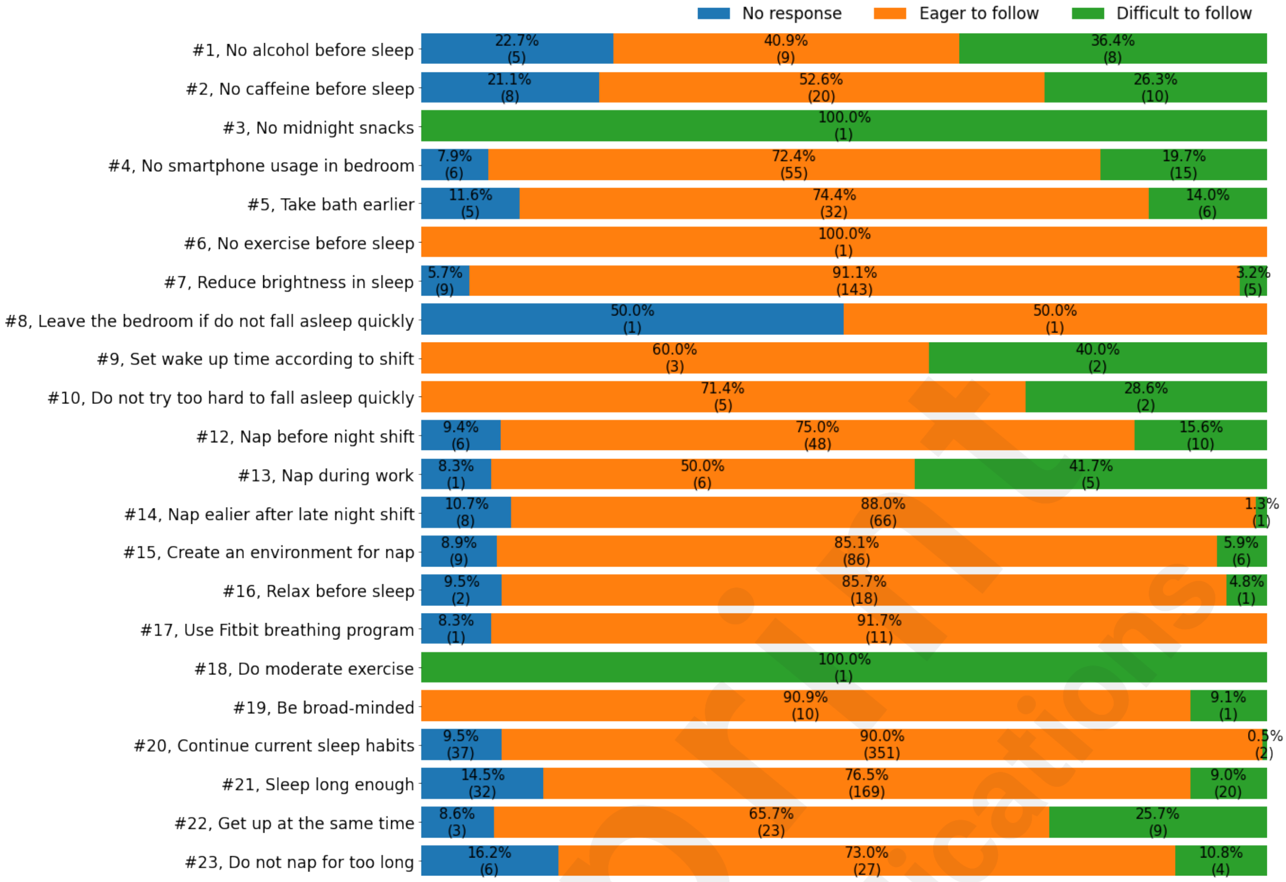

Figure 3: Frequency and proportion of message responses in a stacked bar plot. Each bar includes the frequency and proportion of participants' responses (no response, eager to follow, difficult to follow) to advice pieces containing each message.

Advice Prediction Results

Following the approach described in the Advice Prediction Models of the Methods section, we obtained classification results from RF on the test set under dependent and independent settings. Table 5 summarizes the performance metrics computed from 10 repeated tuning and training of the classifier, 10 random guessing trials, and a classifier that always predicts the positive class. The performance of the RF was compared with random guessing by Welch's t-test and was compared with the positive class classifier by one-sample t-test.

As shown by Table 5, RF achieved higher F1 scores in all but one classification task than random guessing and always predicting the positive class. Among 28 t-tests conducted, performance differences were statistically significant with $p < 0.001$ for 24 tests and $p < 0.05$ for 3 tests. No significant differences were found under independent settings for predictions of message 20 (continue current sleep habits) compared with always positive prediction. The precision and recall scores of the proposed RF are reported in Table S3. RF obtained low precision but high recall scores in prediction tasks: most precision scores are below 0.4, while all recall scores are above 0.6.

Table 5: Advice prediction F1 scores of the proposed RF and two baselines (random guessing and always positive prediction). One-side Welch's t-test was used to compare RF to random guessing, and one-sample t-test to compare RF to always positive prediction.

Setting	Message	Random Forest	Random	Always positive

	ID		guessing	
		F1 mean (SD)	F1 mean (SD; *P*)	F1 mean (*P*)
Dependent				
	4	0.37 (0.03)	0.19 (0.024; <.001)	0.20 (<.001)
	7	0.51 (0.01)	0.25 (0.040; <.001)	0.28 (<.001)
	12	0.36 (0.03)	0.18 (0.032; <.001)	0.19 (<.001)
	14	0.30 (0.01)	0.15 (0.028; <.001)	0.19 (<.001)
	15	0.35 (0.02)	0.24 (0.038; <.001)	0.30 (<.001)
	20	0.73 (0.01)	0.52 (0.026; <.001)	0.71 (<.001)
	21	0.46 (0.04)	0.31 (0.031; <.001)	0.35 (<.001)
Independent				
	4	0.24 (0.01)	0.15 (0.015; <.001)	0.17 (<.001)
	7	0.63 (0.03)	0.26 (0.022; <.001)	0.25 (<.001)
	12	0.09 (0.02)	0.05 (0.015; <.001)	0.06 (*P*=.001)
	14	0.14 (0.02)	0.09 (0.015; <.001)	0.09 (<.001)
	15	0.14 (0.02)	0.10 (0.031; *P*=.001)	0.11 (.002)
	20	0.66 (0.01)	0.51 (0.044; <.001)	0.67 (*P*=.82)
	21	0.42 (0.02)	0.31 (0.029; <.001)	0.38 (<.001)

Important Features

Different important features were found in the RF classifiers for each message label. For some labels, there were one or two dominant features with a much larger decrease in F1 scores than other features after random permutations. Table 6 lists these dominant features for predictions of message 4 (refrain from using a smartphone in the bedroom), 7 (darken the bedroom when you go to bed), 20 (continue current sleep habits), and 21 (try to make enough time for sleep) under dependent and independent settings. The dominant features for messages 4, 7, and 21 matched the messages. For instance, for message 7, the models primarily examined the average brightness of the sleep environment to make predictions, and average sleep duration played a significant role in predicting message 21. On the other hand, no dominant features were found for messages 12 (take a nap for approximately 90 minutes before the night shift), 14 (take a nap earlier after a late-night shift to refresh), and 15 (create an environment for taking a nap), and the highest average F1 score decrease among all features was less than 0.01.

Table 6: Dominant features for predictions of message 4 (refrain from using a smartphone in the bedroom), 7 (darken the bedroom when you go to bed), 20 (continue current sleep habits), and 21 (try to make enough time for sleep) under dependent and independent settings.

Setting	Message ID	Feature	Mean F1 decrease (SD)
Dependent			
	4	Average amount of phone use prior to sleep	0.11 (0.04)
	7	Average brightness of sleep environment	0.38 (0.04)
	20	Response to previous selection of message 20	0.12 (0.04)
		Average fatigue recovery level	0.04 (0.01)
	21	Average sleep duration	0.03 (0.02)
		Response to previous selection of message 20	0.01 (0.02)
Independent			
	4	Average amount of phone use prior to sleep	0.19 (0.04)
	7	Average brightness of sleep environment	0.29 (0.06)
	20	Response to previous selection of message 20	0.15 (0.02)
		Average fatigue recovery level	0.03 (0.01)
	21	Average sleep duration	0.03 (0.02)
		Response to previous selection of message 20	0.08 (0.04)

Discussion

Principal Results

In this study, we used daily surveys and Fitbit to obtain physiological and behavioral data of ICU shift workers, which were reviewed by physicians to provide sleep advice to them. We extracted features from the collected data and (1) conducted analyses to characterize participants' behaviors and advice and (2) constructed and evaluated random forest classifiers to predict sleep advice selections.

Participant Characteristics

Hierarchical clustering discovered different work shifts and behavior patterns among clusters and subclusters, and we investigated their relationships with their wellbeing scores. First, work shifts today contributed to low wellbeing scores for clusters 3 and 5, and several wellbeing metrics were further distinguished by their different shift patterns. For example, cluster 3 had a much shorter average sleep duration of around 220 minutes and the lowest sleep quality. This was potentially caused by day shifts the day before and midnight shifts on the day for almost all samples. In this cluster, participants finished their work at around 4 p.m. the day before, then they started the next shift period around midnight and finished at around 8 a.m. on the following day. After work, they filled out the morning surveys and reported their wellbeing, sleep duration, and sleep quality. Low scores on these metrics reflected their tiredness shortly after work and lack of rest between work shifts.

On the other hand, most samples in cluster 5 did not have shifts the day before but day shifts on the day. As a result, participants could get more time for sleep reflected by a longer average sleep duration of 400 minutes, better sleep quality, and higher wellbeing scores in the morning compared with samples of cluster 3. Such discrepancies in shift patterns and wellbeing metrics can even be observed between subclusters d and k of the same larger cluster 2. Although most samples of both subclusters had midnight shifts the day before, subcluster d majorly did not have shifts on the day and achieved an average level of wellbeing scores. Profile of subcluster k is more similar to that of cluster 5 with midnight shifts today and much worse wellbeing and sleep quality than subcluster d, which is also supported by the t-SNE plot (right plot of Figure S1) that subcluster k is closer to subcluster h (equivalence of cluster 3) than to subcluster d. Furthermore, the majority of cluster 1 does not have a work shift today which leads to its high wellbeing scores as one day off probably helped participants relax. Nevertheless, shifts today are not always negatively correlated with wellbeing. Specifically, cluster 0 has above average wellbeing scores although most samples have work shifts that day.

Furthermore, several subclusters are characterized by some distinct features of their profiles. For example, subclusters i and l of cluster 5 have similar work patterns, but they differ significantly in average daily step counts where subcluster i has the highest count of over 10000 among all subclusters while l has the lowest count of fewer than 2000. Although multiple clusters and subclusters contain samples with overtime shifts longer than 8 hours, subcluster e is the only one with more than 9 hours of average work duration for three consecutive days (10.6 hours the day before yesterday, 9.4 hours yesterday, and 11.5 hours today). The long work durations might also contribute to other characteristics of this subcluster including the shortest time to fall asleep and the highest daily caffeine consumption. Despite the existence of some outliers, the clusters and subclusters generally grouped well as shown by Figure 4, and their profiles summarized from average feature values provided many insights into participants' daily behaviors, shifts, wellbeing, and their interrelationships.

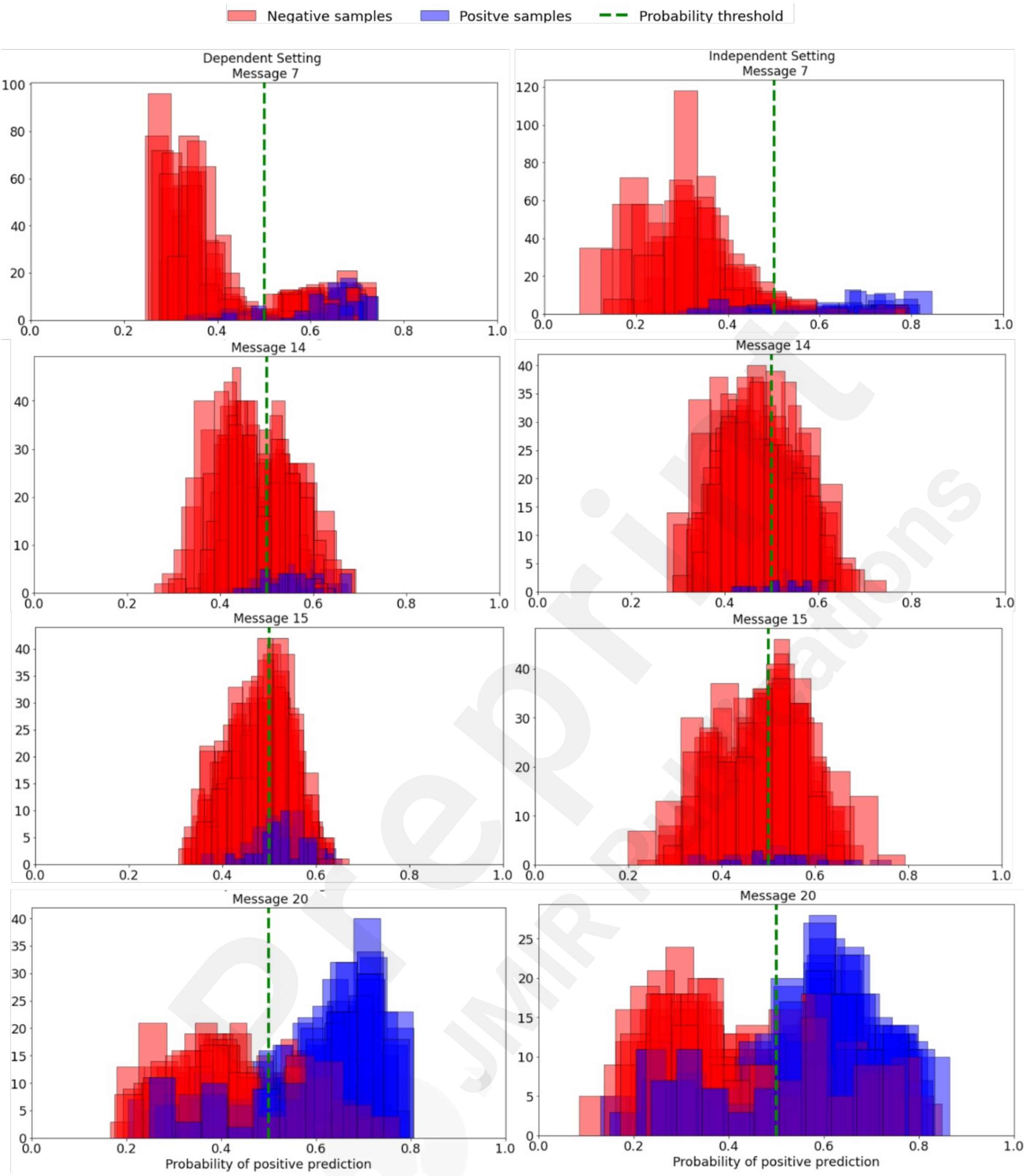

Figure 4: Distributions of advice prediction probabilities for message 7 (darken the bedroom when you go to bed), 14 (take a nap earlier after a late-night shift to refresh), 15 (create an environment for taking a nap), and 20 (continue current sleep habits) under dependent and independent settings. Each graph contains histograms of all 10 repetitions of one advice prediction model where each histogram depicts distributions of positive prediction probabilities for all actual positive samples (blue) and negative samples (red) separately. The green dashed line corresponds to the probability threshold of 0.5 such that samples with probabilities less than the threshold would be predicted as negative and vice versa.

Advice Prediction and Probabilities

Our advice prediction models achieved F1 scores higher than baselines, but the scores are not very high on the absolute scale. Among the 14 models constructed, the highest score achieved is 0.73 and 10 models scored lower than 0.5. Precision scores are also much lower than recall scores, which means there were many false positives although most actual advice selections were predicted. One explanation for low F1 scores is class imbalance. As shown in Advice Statistics in the Results section, only the most frequent message label had a balanced class distribution, while for all the other labels less than 30% of all samples were from the positive class.

Although the training data was balanced by undersampling the negative class, the testing data was not balanced. As a result, the testing data has very high skew values defined as the ratio between the number of negative class samples and the number of positive class samples, which can lead to much lower F1 scores as illustrated by Jeni et al [50]. According to their simulation study, a classifier with a 20% misclassification rate could only yield an F1 score of around 0.4 when skew was 10 which further decreased to 0.2 when skew increased to 30. For our testing data, negative class is the majority class except for message 20 (continue current sleep habits), and skew varies from 0.80 to 8.27 under the user-dependent setting and from 0.99 to 31.43 under the user-independent setting. Following their suggestions, we computed average skew-normalized F1 scores by undersampling the majority class of the testing data for every repetition of each advice prediction model and reported the results in Table 7. After skew normalization, F1 scores increased significantly and ranged from 0.63 to 0.76 for all advice prediction models.

Table 7: Skew of testing set, original average F1 score, and average skew-normalized F1 score of advice prediction models

Setting	Message ID	Skew	Original F1 mean (SD)	Skew-normalized F1 mean (SD)
Dependent				
	4	7.93	0.37 (0.03)	0.76 (0.03)
	7	5.18	0.51 (0.01)	0.75 (0.02)
	12	8.27	0.36 (0.03)	0.75 (0.05)
	14	8.27	0.30 (0.01)	0.72 (0.03)
	15	4.74	0.35 (0.02)	0.63 (0.03)
	20	0.8	0.73 (0.01)	0.72 (0.01)
	21	3.63	0.46 (0.04)	0.67 (0.05)
Independent				
	4	9.81	0.24 (0.01)	0.73 (0.01)
	7	5.88	0.63 (0.03)	0.73 (0.06)
	12	31.43	0.09 (0.02)	0.68 (0.11)
	14	19.64	0.14 (0.02)	0.65 (0.07)
	15	16.46	0.14 (0.02)	0.63 (0.07)
	20	0.99	0.66 (0.01)	0.65 (0.01)
	21	3.28	0.42 (0.02)	0.63 (0.03)

Our advice prediction models predicted not only message selections but also probabilities of the selections. For each repetition of one advice prediction model, we plotted distributions of positive prediction probabilities for all actual positive samples and negative samples separately in histograms and included histograms for all 10 repetitions of the model in one graph. Figure 4 includes distributions of advice prediction probabilities for message 7 (Darken the bedroom when you go to bed.), 14 (Take a nap earlier after a late-night shift to refresh.), 15 (Create an environment for taking a nap.), and 20 (Continue current sleep habits.) under user dependent and independent settings to illustrate typical patterns of the probability distributions. Distributions vary greatly among these models and help explain discrepancies in prediction performance. To be specific, models with left-skewed distributions of positive samples tended to achieve higher recall scores.

Under left-skewed distributions, the majority of positive samples were located to the right of the threshold with positive prediction probabilities higher than 0.5 resulting in relatively fewer false negatives and higher recall scores. Examples of such distributions of positive samples include distributions of messages 7 and 20 as illustrated in Figure 4, as well as distributions of message 4 (refrain from using a smartphone in the bedroom). On the other hand, when distributions of positive samples were approximately bell-shaped, recall scores depended on distribution centers. For example, both positive sample distributions of messages14 and 15 under the user-dependent setting were roughly bell-shaped and centered on the right of the threshold in Figure 4, but distributions of message 14 were slightly more to the right. As a result, the model for message 14 achieved higher recall scores than the model for message 15 since fewer false negatives were produced in the distributions. Under the user-independent setting, the distribution centers of both message labels were close to each other, and the resulting recall scores were similar. Such a relationship between distribution centers and recall scores also existed in messages 21 (try to make enough time for sleep.) and 12 (take a nap for approximately 90 minutes before the night shift).

While recall scores were only related to distributions of positive samples, precision scores were also affected by distributions of negative samples. When distributions of negative samples were about bell-shaped or not right skewed enough and there were more negative samples than positive samples, models tended to yield low precision scores as happened for most advice prediction models because of substantial overlaps between negative samples and positive samples to the right of the threshold. Exceptions to these conditions were models for message 20 under dependent and independent settings and model for message 7 under the user-independent setting. For message 20, although the distributions of negative samples were not extremely right-skewed, there were many more positive samples than negative samples to the right of the threshold since positive sample distributions were left-skewed and classes were approximately balanced. For message 7, despite there being more negative samples than positive samples, distributions of negative samples were greatly right-skewed, which left fewer negative samples to the right of the threshold and resulted in relatively higher precision scores.

Using the histograms of advice prediction probabilities, we evaluated the performance of advice prediction models beyond simple metric values and discovered the relationship between model performance and distributions of prediction probabilities.

Comparison with Prior Work

Our system provided sleep and behavior recommendations for shift workers with poor sleep quality. Although our system did not follow the CBTI protocol, we borrowed elements from it such as sleep hygiene advice and relaxation. We designed sleep recommendations specifically targeted for shift workers including strategic napping and clockwise shift rotation. Our system offered several advantages over existing CBTI and sleep recommender systems. First of all, while most mobile and online CBTI systems required users to review sleep-related learning modules, our system reduced users' commitment by sending them simple pieces of advice to follow, which could be beneficial as shift workers often suffered from poor work-life balance [51]. Consequently, our study yielded a low dropout rate of 5% (3/64), as opposed to rates of other CBTI systems ranging from 4% to 61% [15,16,18–23,26,29,52].

Moreover, our system received the most support from sleep specialists compared to existing sleep recommender systems, which only had limited or even no clinician support. For example, clinicians only provided recommendation templates for further customization by the automated Sleepcoacher system [30]. Cohort-based sleep recommender relied on profiles of similar users as the only recommendation source [31]. The system developed by Pandey et al constructed its recommendations based on discovered relationships between the user's lifestyle and sleep [32]. The lack of clinician support could impair the credibility of provided sleep recommendations among users. For instance, users of Sleepcoacher suggested adding justifications for the advice, and several users of the cohort-based sleep recommender thought the recommendations were implausible and even untrustworthy as the recommendations contradicted their own beliefs [30,31]. Our system, however, only delivered clinician-reviewed recommendations, which greatly enhanced their trustworthiness.

Furthermore, only our study evaluated the accuracy of the algorithm for automatic sleep advice provision. Many existing CBTI systems were able to provide sleep restriction and sleep hygiene recommendations automatically, but no analysis was available to evaluate how close these generated recommendations are to actual clinicians' advice. Similarly, sleep recommender systems did not compare their recommendations with professionals' suggestions. Specifically, clinicians prepared all recommendations in the preliminary study of Sleepcoacher, and the provisions were automated by a correlation-based algorithm in the final study, but the study did not evaluate the accuracy of the algorithm for substituting clinicians [30]. Despite the effectiveness of these systems, an analysis of algorithm accuracy could further justify provided recommendations and resolve users' doubts about their credibility. On the other hand, although our system is not automated at the current stage, we developed an algorithm and assessed its accuracy for automatic advice provision, which showed decent performance and great potential in providing clinical-level sleep recommendations.

Limitations and Future Work

False Predictions

Despite higher F1 scores after skew normalization, false negative and false positive predictions still exist and prevent F1 scores from further improvements. To investigate the possible causes of these incorrect predictions, we examined the relationship between feature values and advice predictions. Specifically, we chose the feature brightness sleep (brightness level of sleep environment) and one repetition of the dependent advice prediction model for message 7 (darken the bedroom when you go to bed). We selected this combination because the feature had a dominant effect on predictions of the message label as shown in the Important Features section and could demonstrate the limitations of the models more clearly with the existence of an explicit decision rule.

For each participant, we plotted the trend of the feature across all days of their enrollment and marked each predicted and the actual selection of the message separately. Figure S2 includes such plots for 6 participants to illustrate incorrect predictions and limitations of the advice prediction models. First, the models did not have memories about previous message selections and might produce false positives as duplicate responses to previous feature values. For example, the model made three positive predictions on days 22nd, 24th, and 26th in response to peak of brightness on the day 22nd for participant 1102, yet only the one on the day 22nd was an actual positive. Similarly, two positive predictions were made for participant 1154 on days 19th and 22nd for a brightness surge on the day 19th but the latter one was a false positive. Shortening the feature averaging window before each message selection might eliminate some of these false positives, but it would increase the risk of producing false negatives when a day with a favorable value for message selection is excluded from the window. Moreover, sometimes a prediction was made as a proper response even though the message was not selected, for instance, the false positive prediction on day 22nd for participant 1154. Although it might be hard to avoid such scenarios, a false positive can probably benefit participants in this case as it suggests a proper response to undesired conditions.

Another scenario of false positives occurred for participants 1315 and 1181. Both participants received 4 false positives although their brightness values only fluctuated within a small range of around 30. A closer look at individual decision trees of the trained random forest showed that some trees considered standard deviations of brightness levels, but most trees made predictions by comparing average brightness with a threshold. Thus, false positive predictions were made because average brightness exceeded the thresholds of most decision trees despite the lack of brightness peaks. However, it could be tricky to make correct predictions under such a scenario. For example, physicians selected the message twice on days 9th and 16th for participant 1315 although brightness remained stable. A plausible explanation could be that physicians initially suggested the participant lower the brightness because it was constantly above the threshold, but later they stopped as neither lower levels nor high peaks of brightness were observed after the suggestions. Nevertheless, it would be very difficult for the model to capture such change.

Unlike false positives, most false negatives occurred only under one scenario: when brightness peaks were not high enough and adjacent days had very low feature values which caused averages to fall below the threshold. For example, on day 33rd for participant 1164 and on days 29th, 31st, and 33rd for participant 1167. Measuring the relative height of a peak (difference between maximum and minimum values of the feature averaging window) instead of averaging might be helpful to eliminate these false negatives, but it may produce other false negatives when feature values remain high and stable. From illustrative plots of feature values and advice predictions, we demonstrated several causes of false positive and false negative predictions, but further investigations and experiments were necessary to properly improve the advice prediction models since some false positives might benefit participants, and strategies to eliminate certain false predictions might introduce other ones.

System Usability

Although our developed algorithm achieved decent performance in predicting sleep advice under both dependent and independent settings, the algorithm is limited to the 7 most popular ones among 23 advice choices due to rare occurrences of the remaining options. Therefore, further studies are necessary to gather enough data to enable predictions for less frequent advice labels.

We obtained clinician’s experiences in using the sleep advice models (N=2) through questionnaires. They commented that they decided their suggestions mainly based on morning survey and work shift and wearable data and after they selected the suggestions, they referred to the output of the models and made the final decisions.

Future studies can compare the intervention effects of a pure clinician input system, an algorithm-assisted system, and a fully automated system to demonstrate the usefulness of the machine learning algorithm.

Given the effectiveness of existing sleep recommender systems with limited clinical support, we believe our automated system has the advantage that it was built upon the knowledge of sleep specialists, even if our algorithm still requires improvement to further boost performance. We chose to evaluate our models in user-dependent and user-independent settings instead of leave-one-out cross-validation (LOOCV) because the advice frequencies were extremely imbalanced among participants. However, such a scheme will fit real-world scenarios the best when every time a new user joins and models keep updating with all available data from other users. It is also worth experimenting with personalization beyond LOOCV by continuously incorporating new user data as they progress to better adapt models to specific user characteristics. Furthermore, similar to some existing research [25,27,28,30,31], it can be beneficial to conduct a usability study on clinicians and shift workers to gather feedback and improve our system.

Conclusions

We developed an online system to provide individualized sleep and behavior advice with CBTI elements for shift workers. We collected data from shift workers in ICU at two hospitals in Japan (N=61) for 5 weeks including Fitbit, survey, sleep advice data. We found clusters of shift workers' physiological and behavioral features using hierarchical clustering and developed and evaluated random forest machine learning algorithms that predicts physicians' sleep advice for shift workers using wearable and survey data. Our analysis showed clusters can be characterized with work shifts and behaviors patterns. For example, days with low sleep duration and the lowest sleep quality was associated with a day shift on the day before and a midnight shift on the current day. Our advice prediction models achieved higher F1 scores compared to the baseline. The analysis of the feature importance of our models showed that the most important features matched the message sent to participants. Although our current system requires physician input, an accurate machine learning algorithm would be promising for automating without hurting the trustworthiness of selected recommendations. The algorithms developed in this study are limited to the 7 most popular ones among 23 choices due to rare occurrences of the remaining options. Therefore, further studies are necessary to gather enough data to enable predictions for less frequent advice labels.

Acknowledgements

This research was funded by the Japan Agency for Medical Research and Development (#19217687) and the National Science Foundation (#1840167).

Abbreviations

CBTI: cognitive behavioral treatment of insomnia
ICU: intense care unit
LOOCV: leave-one-out cross-validation
RF: random forest
SD: standard deviation
SRI: sleep regularity index
SWSD: shift work sleep disorder
t-SNE: t-distributed stochastic neighbor embedding

References

1. Alterman T, Luckhaupt SE, Dahlhamer JM, Ward BW, Calvert GM. Prevalence rates of work organization characteristics among workers in the U.S.: Data from the 2010 National Health Interview Survey. Am J Ind Med 2013 Jun;56(6):647–659. doi: 10.1002/ajim.22108
2. Wickwire EM, Geiger-Brown J, Scharf SM, Drake CL. Shift Work and Shift Work Sleep Disorder. Chest 2017 May;151(5):1156–1172. doi: 10.1016/j.chest.2016.12.007
3. Haile KK, Asnakew S, Waja T, Kerbih HB. Shift work sleep disorders and associated factors among nurses at federal government hospitals in Ethiopia: a cross-sectional study. BMJ Open 2019 Aug 27;9(8):e029802. doi: 10.1136/bmjopen-2019-029802
4. Flo E, Pallesen S, Magerøy N, Moen BE, Grønli J, Hilde Nordhus I, Bjorvatn B. Shift Work Disorder in Nurses – Assessment, Prevalence and Related Health Problems. PLoS One 2012 Apr 2;7(4):e33981. doi: 10.1371/journal.pone.0033981

5. Di Milia L, Waage S, Pallesen S, Bjorvatn B. Shift Work Disorder in a Random Population Sample – Prevalence and Comorbidities. PLoS One 2013 Jan 25;8(1):e55306. doi: 10.1371/journal.pone.0055306
6. Cheng W-J, Cheng Y. Night shift and rotating shift in association with sleep problems, burnout and minor mental disorder in male and female employees. Occup Environ Med 2017 Jul;74(7):483–488. doi: 10.1136/oemed-2016-103898
7. Khan WAA, Jackson ML, Kennedy GA, Conduit R. A field investigation of the relationship between rotating shifts, sleep, mental health and physical activity of Australian paramedics. Sci Rep 2021 Jan 13;11(1):866. doi: 10.1038/s41598-020-79093-5
8. Saleh AM, Awadalla NJ, El-masri YM, Sleem WF. Impacts of nurses’ circadian rhythm sleep disorders, fatigue, and depression on medication administration errors. Egyptian Journal of Chest Diseases and Tuberculosis 2014 Jan;63(1):145–153. doi: 10.1016/j.ejcdt.2013.10.001
9. Blytt KM, Bjorvatn B, Moen BE, Pallesen S, Harris A, Waage S. The association between shift work disorder and turnover intention among nurses. BMC Nurs 2022 Dec 6;21(1):143. doi: 10.1186/s12912-022-00928-9
10. Wright KP, Bogan RK, Wyatt JK. Shift work and the assessment and management of shift work disorder (SWD). Sleep Med Rev 2013 Feb;17(1):41–54. doi: 10.1016/j.smrv.2012.02.002
11. Qaseem A, Kansagara D, Forciea MA, Cooke M, Denberg TD. Management of Chronic Insomnia Disorder in Adults: A Clinical Practice Guideline From the American College of Physicians. Ann Intern Med 2016 Jul 19;165(2):125. doi: 10.7326/M15-2175
12. Williams J, Roth A, Vatthauer K, McCrae CS. Cognitive Behavioral Treatment of Insomnia. Chest 2013 Feb;143(2):554–565. doi: 10.1378/chest.12-0731
13. Järnefelt H, Lagerstedt R, Kajaste S, Sallinen M, Savolainen A, Hublin C. Cognitive behavioral therapy for shift workers with chronic insomnia. Sleep Med 2012 Dec;13(10):1238–1246. doi: 10.1016/j.sleep.2012.10.003
14. Järnefelt H, Sallinen M, Luukkonen R, Kajaste S, Savolainen A, Hublin C. Cognitive behavioral therapy for chronic insomnia in occupational health services: Analyses of outcomes up to 24 months post-treatment. Behaviour Research and Therapy 2014 May;56:16–21. doi: 10.1016/ j.brat.2014.02.007
15. Järnefelt H, Härmä M, Sallinen M, Virkkala J, Paajanen T, Martimo K-P, Hublin C. Cognitive behavioural therapy interventions for insomnia among shift workers: RCT in an occupational health setting. Int Arch Occup Environ Health 2020 Jul 18;93(5):535–550. doi: 10.1007/s00420-019-01504-6
16. Peter L, Reindl R, Zauter S, Hillemacher T, Richter K. Effectiveness of an Online CBT-I Intervention and a Face-to-Face Treatment for Shift Work Sleep Disorder: A Comparison of Sleep Diary Data. Int J Environ Res Public Health 2019 Aug 24;16(17):3081. doi: 10.3390/ijerph16173081
17. Retzer L, Feil M, Reindl R, Richter K, Lehmann R, Stemmler M, Graessel E. Anonymous online cognitive behavioral therapy for sleep disorders in shift workers—a study protocol for a randomized controlled trial. Trials 2021 Dec 16;22(1):539. doi: 10.1186/s13063-021-05437-9
18. Ritterband LM, Thorndike FP, Gonder-Frederick LA, Magee JC, Bailey ET, Saylor DK, Morin CM. Efficacy of an Internet-Based Behavioral Intervention for Adults With Insomnia. Arch Gen Psychiatry 2009 Jul 1;66(7):692. doi: 10.1001/archgenpsychiatry.2009.66
19. Espie CA, Kyle SD, Williams C, Ong JC, Douglas NJ, Hames P, Brown JSL. A Randomized, Placebo-Controlled Trial of Online Cognitive Behavioral Therapy for Chronic Insomnia Disorder Delivered via an Automated Media-Rich Web Application. Sleep 2012 Jun;35(6):769–781. doi: 10.5665/sleep.1872
20. Batterham PJ, Christensen H, Mackinnon AJ, Gosling JA, Thorndike FP, Ritterband LM, Glozier N, Griffiths KM. Trajectories of change and long-term outcomes in a randomised controlled trial of internet-based insomnia treatment to prevent depression. BJPsych Open 2017 Sep;3(5):228–235. PMID:28959453

21. Hagatun S, Vedaa Ø, Nordgreen T, Smith ORF, Pallesen S, Havik OE, Bjorvatn B, Thorndike FP, Ritterband LM, Sivertsen B. The Short-Term Efficacy of an Unguided Internet-Based Cognitive-Behavioral Therapy for Insomnia: A Randomized Controlled Trial With a Six-Month Nonrandomized Follow-Up. Behavioral Sleep Medicine 2019 Mar 4;17(2):137–155. doi: 10.1080/15402002.2017.1301941
22. Cheng P, Luik AI, Fellman-Couture C, Peterson E, Joseph CLM, Tallent G, Tran KM, Ahmedani BK, Roehrs T, Roth T, Drake CL. Efficacy of digital CBT for insomnia to reduce depression across demographic groups: a randomized trial. Psychol Med 2019 Feb 24;49(3):491–500. doi: 10.1017/S0033291718001113
23. Freeman D, Sheaves B, Goodwin GM, Yu L-M, Nickless A, Harrison PJ, Emsley R, Luik AI, Foster RG, Wadekar V, Hinds C, Gumley A, Jones R, Lightman S, Jones S, Bentall R, Kinderman P, Rowse G, Brugha T, Blagrove M, Gregory AM, Fleming L, Walklet E, Glazebrook C, Davies EB, Hollis C, Haddock G, John B, Coulson M, Fowler D, Pugh K, Cape J, Moseley P, Brown G, Hughes C, Obonsawin M, Coker S, Watkins E, Schwannauer M, MacMahon K, Siriwardena AN, Espie CA. The effects of improving sleep on mental health (OASIS): a randomised controlled trial with mediation analysis. Lancet Psychiatry 2017 Oct;4(10):749–758. doi: 10.1016/S2215-0366(17)30328-0
24. Beun RJ, Brinkman W-P, Fitrianie S, Griffioen-Both F, Horsch C, Lancee J, Spruit S. Improving Adherence in Automated e-Coaching. 2016. p. 276–287. doi: 10.1007/978-3-319-31510-2_24
25. Beun RJ, Fitrianie S, Griffioen-Both F, Spruit S, Horsch C, Lancee J, Brinkman W-P. Talk and Tools: the best of both worlds in mobile user interfaces for E-coaching. Pers Ubiquitous Comput 2017 Aug 19;21(4):661–674. doi: 10.1007/s00779-017-1021-5
26. Horsch CH, Lancee J, Griffioen-Both F, Spruit S, Fitrianie S, Neerincx MA, Beun RJ, Brinkman W-P. Mobile Phone-Delivered Cognitive Behavioral Therapy for Insomnia: A Randomized Waitlist Controlled Trial. J Med Internet Res 2017 Apr 11;19(4):e70. doi: 10.2196/jmir.6524
27. Kuhn E, Weiss BJ, Taylor KL, Hoffman JE, Ramsey KM, Manber R, Gehrman P, Crowley JJ, Ruzek JI, Trockel M. CBT-I Coach: A Description and Clinician Perceptions of a Mobile App for Cognitive Behavioral Therapy for Insomnia. Journal of Clinical Sleep Medicine 2016 Apr 15;12(04):597–606. doi: 10.5664/jcsm.5700
28. Koffel E, Kuhn E, Petsoulis N, Erbes CR, Anders S, Hoffman JE, Ruzek JI, Polusny MA. A randomized controlled pilot study of CBT-I Coach: Feasibility, acceptability, and potential impact of a mobile phone application for patients in cognitive behavioral therapy for insomnia. Health Informatics J 2018 Mar 27;24(1):3–13. doi: 10.1177/1460458216656472
29. Kuhn E, Miller KE, Puran D, Wielgosz J, YorkWilliams SL, Owen JE, Jaworski BK, Hallenbeck HW, McCaslin SE, Taylor KL. A Pilot Randomized Controlled Trial of the Insomnia Coach Mobile App to Assess Its Feasibility, Acceptability, and Potential Efficacy. Behav Ther 2022 May;53(3):440–457. doi: 10.1016/j.beth.2021.11.003
30. Daskalova N, Metaxa-Kakavouli D, Tran A, Nugent N, Boergers J, McGeary J, Huang J. SleepCoacher. Proceedings of the 29th Annual Symposium on User Interface Software and Technology New York, NY, USA: ACM; 2016. p. 347–358. doi: 10.1145/2984511.2984534
31. Daskalova N, Lee B, Huang J, Ni C, Lundin J. Investigating the Effectiveness of Cohort-Based Sleep Recommendations. Proc ACM Interact Mob Wearable Ubiquitous Technol New York, NY, USA: Association for Computing Machinery; 2018 Sep;2(3). doi: 10.1145/3264911
32. Pandey V, Upadhyay D, Nag N, Jain R. Personalized User Modelling for Context-Aware Lifestyle Recommendations to Improve Sleep. 2020.
33. Kay M, Choe EK, Shepherd J, Greenstein B, Watson N, Consolvo S, Kientz JA. Lullaby. Proceedings of the 2012 ACM Conference on Ubiquitous Computing New York, NY, USA: ACM; 2012. p. 226–234. doi: 10.1145/2370216.2370253
34. Bauer JS, Consolvo S, Greenstein B, Schooler J, Wu E, Watson NF, Kientz J. ShutEye. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems New York, NY, USA: ACM; 2012. p. 1401–1410. doi: 10.1145/2207676.2208600

35. Daskalova N, Yoon J, Wang Y, Araujo C, Beltran G, Nugent N, McGeary J, Williams JJ, Huang J. SleepBandits: Guided Flexible Self-Experiments for Sleep. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems New York, NY, USA: ACM; 2020. p. 1–13. doi: 10.1145/3313831.3376584
36. Ito-Masui A, Kawamoto E, Sakamoto R, Yu H, Sano A, Motomura E, Tanii H, Sakano S, Esumi R, Imai H, Shimaoka M. Internet-Based Individualized Cognitive Behavioral Therapy for Shift Work Sleep Disorder Empowered by Well-Being Prediction: Protocol for a Pilot Study. JMIR Res Protoc 2021;10(3):e24799. doi: 10.2196/24799
37. Yu H, Itoh A, Sakamoto R, Shimaoka M, Sano A. Forecasting Health and Wellbeing for Shift Workers Using Job-Role Based Deep Neural Network. 2021. p. 89–103. doi: 10.1007/978-3-030-70569-5_6
38. Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. American Journal of Physiology-Heart and Circulatory Physiology 2000 Jun 1;278(6):H2039–H2049. doi: 10.1152/ajpheart.2000.278.6.H2039
39. Phillips AJK, Clerx WM, O'Brien CS, Sano A, Barger LK, Picard RW, Lockley SW, Klerman EB, Czeisler CA. Irregular sleep/wake patterns are associated with poorer academic performance and delayed circadian and sleep/wake timing. Sci Rep 2017 Jun 12;7(1):3216. doi: 10.1038/s41598-017-03171-4
40. Lunsford-Avery JR, Engelhard MM, Navar AM, Kollins SH. Validation of the Sleep Regularity Index in Older Adults and Associations with Cardiometabolic Risk. Sci Rep 2018 Sep 21;8(1):14158. doi: 10.1038/s41598-018-32402-5
41. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J Classif 2014 Oct 18;31(3):274–295. doi: 10.1007/s00357-014-9161-z
42. Nielsen F. Hierarchical Clustering. Introduction to HPC with MPI for Data Science Cham: Springer International Publishing; 2016. p. 195–211. doi: 10.1007/978-3-319-21903-5_8
43. van der Maaten L, Hinton G. Viualizing data using t-SNE. Journal of Machine Learning Research 2008 Mar;9:2579–2605.
44. Fawagreh K, Gaber MM, Elyan E. Random forests: from early developments to recent advancements. Systems Science & Control Engineering 2014 Dec 6;2(1):602–609. doi: 10.1080/21642583.2014.956265
45. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. 2016 Sep 21;
46. Chen C, Breiman L. Using Random Forest to Learn Imbalanced Data. University of California, Berkeley 2004 Mar;
47. Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. J Mach Learn Res JMLR.org; 2012 Feb;13(null):281–305.
48. Breiman L. Random Forests. Mach Learn 2001;45(1):5–32. doi: 10.1023/A:1010933404324
49. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 1991;21(3):660–674. doi: 10.1109/21.97458
50. Jeni LA, Cohn JF, De La Torre F. Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction IEEE; 2013. p. 245–251. doi: 10.1109/ACII.2013.47
51. Puttonen S, Härmä M, Hublin C. Shift work and cardiovascular disease - Pathways from circadian stress to morbidity. Scand J Work Environ Health 2010 Mar;36:96–108. doi: 10.2307/40967836
52. Ritterband LM, Thorndike FP, Ingersoll KS, Lord HR, Gonder-Frederick L, Frederick C, Quigg MS, Cohn WF, Morin CM. Effect of a Web-Based Cognitive Behavior Therapy for Insomnia Intervention With 1-Year Follow-up. JAMA Psychiatry 2017 Jan 1;74(1):68. doi: 10.1001/jamapsychiatry.2016.3249

Supplementary Files