Learning Sleep Quality from Daily Logs

Sungkyu Park
Graduate School of Culture Technology, KAIST
shaun.park@kaist.ac.kr

Cheng-Te Li
Institute of Data Science, National Cheng Kung University
chengte@mail.ncku.edu.tw

Sungwon Han
School of Computing, KAIST
lion4151@kaist.ac.kr

Cheng Hsu
Institute of Data Science, National Cheng Kung University
hsuchengmath@gmail.com

Sang Won Lee
Kyungpook National University Chilgok Hospital, South Korea
leesangwon.psy@gmail.com

Meeyoung Cha
[1]Data Science Group, IBS
[2]School of Computing, KAIST
meeyoung.cha@gmail.com

ABSTRACT

Precision psychiatry is a new research field that uses advanced data mining over a wide range of neural, behavioral, psychological, and physiological data sources for classification of mental health conditions. This study presents a computational framework for predicting sleep efficiency of insomnia sufferers. A smart band experiment is conducted to collect heterogeneous data, including sleep records, daily activities, and demographics, whose missing values are imputed via Improved Generative Adversarial Imputation Networks (Imp-GAIN). Equipped with the imputed data, we predict sleep efficiency of individual users with a proposed interpretable LSTM-Attention (LA Block) neural network model. We also propose a model, Pairwise Learning-based Ranking Generation (PLRG), to rank users with high insomnia potential in the next day. We discuss implications of our findings from the perspective of a psychiatric practitioner. Our computational framework can be used for other applications that analyze and handle noisy and incomplete time-series human activity data in the domain of precision psychiatry.

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**; **Health informatics**;

KEYWORDS

Insomnia; Precision Psychiatry; Data Imputation; Time-series Data; Interpretability; Ranking Model

ACM Reference Format:
Sungkyu Park, Cheng-Te Li, Sungwon Han, Cheng Hsu, Sang Won Lee, and Meeyoung Cha. 2019. Learning Sleep Quality from Daily Logs. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330792

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330792

1 INTRODUCTION

Insomnia is a common psychiatric illness that can decrease quality of life.[1] Approximately 30 percent of people from different countries report to suffer from one or more of the symptoms of insomnia [26]. Research has shown that various factors including sleep irregularity, excessive coffee or alcohol, lack of physical activity, and nap can aggravate symptoms of insomnia. Therefore, insomnia sufferers may each exhibit different behavioral characteristics even though their exposed symptoms might be similar.

Recently, precision psychiatry has emerged as a new concept. In precision psychiatry, individual variability, such as the genetic information, neural circuits, individual characteristics, medical codes and electronic health records (EHR) are carefully considered to collectively arrive at a diagnosis, treatment plan, and prediction of prognosis [10, 15]. Its premise is that jointly analyzing heterogeneous data sources can yield more accurate classification of major psychiatric illnesses than manual classification like the Diagnostic and Statistical Manual of Mental Disorders (DSM) and International Classification of Diseases (ICD) [8].

For insomnia sufferers, different biological and behavioral characteristics can affect the maintenance of symptoms. While genetic and neuroimaging data are costly to collect, behavioral and sleep records are becoming accessible. Despite the potential, less effort has been paid to utilizing the everyday behavioral information for helping treat high-risk patients regarding sleep disorders, which is the main contribution of this work. This work demonstrates how data gathered from popular wearable devices like smart bands can be used to learn insomnia patterns and predict sleep quality. We conducted a 6-week long experiment with 50 participants wearing Fitbit[2] to gather time series data describing both daily sleep and behavioral records. In utilizing heterogeneous data sources, a challenge arises due to data consistency because human wearable devices are likely subject to various noise and biases largely due to missing data (e.g., people forgetting to wear devices or battery runs out) [19]. Therefore, prior to building a holistic model of sleep quality, we treat data consistency problem by proposing an advanced technique called Improved Generative Adversarial Imputation Network. We then develop two use cases in learning

[1]Insomnia is defined by the presence of an individual's report of difficulty with sleep. This may involve clinical diagnosis as well as self-diagnosis based on a general question, "Do you have difficulty falling or staying asleep?"
[2]A smart band that tracks daily activities and sleep behaviors. www.fitbit.com/charge2

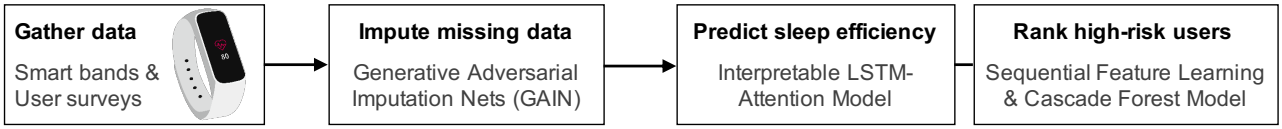


Figure 1: Overall architecture of this paper. We collect data via smart bands (Fitbit Charge 2), followed by missing data imputation by an improved GAIN model. Then we predict future sleep efficiency of users via a proposed interpretable LSTM-Attention model. We also find future high insomnia users via ranking prediction with sequential and interaction feature learning.

sleep quality as below (Fig. 1). Note that codes developed for this work can be accessed via GitHub.[3]

- **Predicting and interpreting sleep efficiency.** Predicting sleep efficiency is one of the crucial tasks in determining the quality of sleep, which can benefit both clinicians and insomnia suffers [7]. In such prediction, it is important to find the root cause of any sleep related dysfunction. Since tonight's sleep is affected by previous day(s)' sleep and activities [3, 16], an interpretable model of sleep efficiency could help develop interventions that are more appropriate in precision psychiatry. The model we present, LSTM-Attention Model, can be used by psychiatrists to diagnose personalized interventions and treatments for insomnia sufferers.
- **Ranking users with high insomnia risk.** Another important problem that we address is ranking users with high insomnia potential in the next day. Compared to estimating the absolute sleep efficiency value itself, a ranking model would be useful for individuals to manage day-to-day symptoms [22, 23]. The model we present, Pairwise Learning-based Ranking Generation (PLRG), can effectively rank users so that individuals with high insomnia risk can be identified.

Several previous research has also predicted sleep efficiency from wearable device data. One study [20] used a k-nearest neighbor classifier to predict changes in sleep-related features while other studies [28] used machine- and deep-learning techniques to classify sleep quality. These studies mainly concentrated on classification accuracy and could not gain enough interpretability due to black-box models. Although a recent study leveraged deep learning techniques to analyze sleep behaviors [27], their aim was at predicting sleep stages based on human polysomnogram signal data that is costly and hard to access.

Contributions. We have developed a deep-learning model that achieves higher prediction performance as well as detailed interpretation compared to other explainable machine learning models, with well generated originally missed data. We successfully predicted not only the absolute value of sleep efficiency (by a proposed LSTM-Attention Ensemble model) but also the ranks of high insomnia-risk users (by a proposed Pairwise Learning-based Ranking Generation model), thereby providing a tool in routinely managing insomnia patients in a robust way. While outperforming existing baselines and state-of-the-arts, our model is also interpretable and brings two major insights: (a) insomnia sufferers mainly exhibit two different periodic rhythms of sleeping, 5 days- and weekly-basis, and (b) both most recent 1-2 days and periodic sleep habits can significantly shape the predictability of sleep efficiency.

[3]https://github.com/Sungwon-Han/Learning-Sleep-Quality-from-Daily-Logs

Table 1: List of data gathered from the smart bands.

Feature	Description	Mean±SD
Modality 1: Sleep Behavior *(Source: Fitbit)*		
sleep_start_time	Time when a user goes to bed	02:45:55±149.4 min
sleep_end_time	Time when a user gets out of bed	09:48:28±156.3 min
onbed_min	Total time duration of staying on bed	422.55±126.55 min
sleep_min	Total time duration of actual sleep	368.78±111.51 min
sleep_efficiency	*sleep_min / onbed_min*	0.87±0.048
awaken_min	Total time duration of wake ups during sleep	53.59±26.41 min
awaken_moments	Total frequency of waking up during sleep	24.08±13.13
nap_min_per_day	Total time duration of naps per day	24.08±55.81 min
nap_freq_per_day	Total frequency of naps per day	0.20±0.44
Modality 2: Daily Activity *(Fitbit)*		
calorie_consume	Total calories consumed per day	2339.3±618.7 kcal
active_calorie	Total calories consumed from activities	993.5±547.3 kcal
walks	Total footstep counts per day	9889±5013 steps
distance	Total distance a user moves per day	6.92±3.56 km
stairs	Total frequency a user takes per day	13.08±12.04
active_ratio	Total moving time wearing a device	0.24±0.11
Modality 3: Personal Demographic *(Survey)*		
age	Age of the participant	23±2.75 (18~29)
gender	Gender information of the participant	Female: 54.76%
BMI	$weight / height^2$	21.9±3.3 (15.9~29.3)
ISI	Insomnia severity index (ISI) score	18.38±2.20 (15~23)

While pursuing each analysis we continuously consulted a psychiatrist (the fifth author) to maintain clinical relevance. Steps that require expert's perspectives include selecting target variables related to sleep, utilizing heterogeneous behavioral logs to track sleep, predicting future sleep efficiency through past records, and interpreting the attention results. We tried to conduct clinically-relevant analyses to gain meaningful insights to help insomnia sufferers.

2 EXPERIMENTAL SETUP

This study utilizes experimental data gathered from wearable fitness devices. We recruit participants via posting a call on an online community of a large university. Since our study focus is insomnia, we screened recruits based on the test result of the Insomnia Severity Index (ISI), which is a brief instrument designed to assess the severity of both nighttime and daytime components of insomnia. A total of 145 students participated in the initial survey, out of whom we invited 50 participants whose ISI scores were 15 or above (i.e., indication of mild to a severe level of insomnia). 45 participants were moderate insomnia sufferers and 5 were severe suffers based on the ISI score. The recruited participants were a balanced group in terms of gender (male : female = 26 : 24) and had similar ages (average age = 23.24 ± 2.92). Besides the age, gender, and the ISI index, subjects were asked to report their heights and weights to calculate the body mass index (BMI). The experiment had been approved by the institutional review board at the authors' institute.

Subjects participated in an experiment wearing a smart band. Fitbit Charge 2 model was provided that logged daily activity and sleep behaviors throughout the experiment period 24/7. The experiment span a total of six weeks from April 23 to June 3, 2018, and the logs were sent to a server daily via a mobile application implemented for this research. The app sent weekly reminders encouraging subjects to keep the devices charged and on. However, seven subjects failed to do so for over two consecutive days and one subject lost the device. We therefore removed their data and used data from the remaining 42 participants (male:female=19:23, average age = 23.00 ± 2.75, 39 moderate sufferers and 3 severe sufferers).

Fitbit reports rich information about one's sleep. For example, it tracks the time when a person lied down (on bed) as well as the predicted time when a person fell asleep. However, Fitbit warns a possible false report of sleep when a person is not moving but neither is asleep for long periods of time. Nonetheless, compared to conventional methods like the Pittsburgh Sleep Quality Index (PSQI) that measures sleep quality and quantity based on self-reported surveys (e.g., when one goes to bed, how long it takes to fall asleep), wearable devices can produce more reliable logs on a daily basis.

Table 1 displays the data features gathered. Fitbit reports activity logs such as the calories consumed, steps walked, total distance moved, stairs, and etc. While Fitbit logs other data like heart rates, they are not accessible by API. Three sleep features (*REM_sleep*, *narrow_sleep*, *deep_sleep*) were excluded from analysis, as they are known to be significantly less accurate than brain (e.g., electroencephalogram, EEG) monitoring [1, 14]. Instead, *sleep_efficiency* was newly added as a feature after consulting a psychiatrist, which is calculated as the fraction of time dedicated to actual sleep out of the time spent lying on bed. A psychiatrist also suggested to utilize two nap-related features including *nap_min_per_day* and *nap_freq_per_day*, because taking a nap can also affect that night's sleep conditions. We also introduced a new variable *active_ratio* that examines the total fraction of time on any activity (i.e., *little*, *very*, or *super*) out of all. Consequently, we could get total 24,864 data points (42 users×42 days×14 behavioral features + 42 users×4 demographic features) while the proportion of missing data was 2.83%.

The experiment subjects, who all suffer from different degree of insomnia, showed diverse sleep patterns. Fig. 2 is sample pattern indicating late sleep hours and irregular naps. It is of great interest to both insomnia sufferers and clinicians to predict in advance any onset of severe insomnia. Offline interventions are based on self-reported sleep logs [5], and the rich context of sleep activities has not been utilized. Data from wearable devices allow clinicians to investigate the fine-grained sleep activities of insomnia sufferers. The core question then remains, which sleep-awake pattern will lead to severe insomnia potential in the near future (i.e., the next day). We examine the time-series data and consider the entire sequence of activities leading to each sleep event to understand the nature of sleep, as suggested in [12].

3 MISSING DATA IMPUTATION

We aim at imputing the collected Fitbit data so that the missing values can be correctly recovered. As mentioned earlier, missing data is one of the primary challenges in utilizing data from human wearable devices. Given there are n users and d data fields, let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{n}$ be

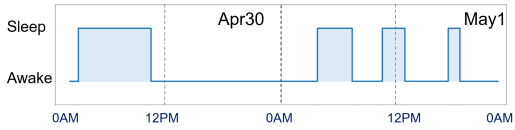


Figure 2: Two-day sleep record of an insomnia sufferer indicating an irregular sleep-awake pattern of Participant #10.

the data set with d features. At a pre-processing step, one can use a certain input transformation function h_j to normalize all values of each feature j ($1 \le j \le d$). It can be represented by $\tilde{\mathbf{X}} = \{\tilde{x}_{ij}\}$, where $\tilde{x}_{ij} = h_j(x_{ij})$. The missing and known values are specified by the following sets: $\mathcal{M} = \{(i,j) : \text{ entry } x_{ij} \text{ is missing}, 1 \le j \le d\}$ and $\mathcal{N} = \{(i,j) : \text{ entry } x_{ij} \text{ is known}, 1 \le j \le d\}$. The general objective of missing data imputation here is to find a real-valued generative function f to impute all missing values $x_{ij} \in \mathcal{M}$ so that each imputed value $\bar{x}_{ij} = f(x_{ij})$ are close to the corresponding realistic but hidden value $\hat{x}_{ij}$. Our goal is to find f that minimizes the mean absolute error (MAE): $\min_f \sum_{(i,j)\in\mathcal{M}} \left|(\hat{x}_{ij} - \bar{x}_{ij})\right| / \|\mathcal{M}\|$.

We leverage the state-of-the-art missing data imputation method, *Generative Adversarial Imputation Nets* (GAIN) [30], to implement the generative function f. GAIN imputes missing values based on the well-known Generative Adversarial Network (GAN) [17]. The reason that we choose GAIN is two-fold. One is that GAIN had been validated to outperform several strong missing data imputation methods, such as MICE, MissForest and AutoEncoder, in a variety of real benchmark datasets. The other is that GAIN is robust to various missing rates, the number of samples, and the feature dimensionality. Despite the powerful imputation capability of GAIN, we still find some unreasonable imputation results. GAIN could generate: (a) *sleep_end_time* earlier than *sleep_start_time*. (b) negative imputed values, which are unreasonable for all of our features. (c) too large or too small feature values that are supposed to have reasonable ranges of values.

To deal with these issues, we propose to perform *post-processing* for the imputed values from GAIN. For each feature j, we define a transformation function g_j to rescale the imputed values. The rescaling function g_j is devised to *shift* and *scale* the range of imputed values such that they fall into the maximum and minimum values of each feature j in the observed data. Such action is to not only ensure no negative imputed values, but also make the imputed values follow the reasonable distribution as the observed data. In addition, it is generally believed that a proper *pre-processing* for the input observed data can also benefit the imputation. Hence, we define the input transformation function h_j using minimum-maximum rescaling. The final imputed value $\bar{x}_{ij}$ of input x_{ij} can be represented by $\bar{x}_{ij} = g_j(f(h_j(x_{ij})))$.

Evaluation. To examine the effectiveness of our improved GAIN method, we randomly split the original non-missing data records into 80% observed (training) data and 20% missing (testing) data so that we have ground truth for evaluation. 5-fold cross validation is conducted. We use *Mean Absolute Error* (MAE) of original feature values as the metric. Several competing methods are employed, including the original GAIN (Ori-GAIN) without input and output transformation, User Average (User-Avg) (averaging all of previous values per feature for each user), and k-Nearest Neighbor (KNN).

Table 2: Results in MAE by various imputation methods.

	Imp-GAIN	Ori-GAIN	User-Avg	KNN
sleep_start_time	**168.3184**	211.1669	172.8133	249.8450
sleep_end_time	**167.1514**	256.4629	226.6629	375.5200
sleep_min	**80.0195**	109.7301	102.6908	137.9300
sleep_efficiency	0.0361	0.0525	**0.0352**	0.0419
awaken_min	**20.9912**	29.0661	20.0440	26.7100
awaken_moments	**7.6124**	14.5176	11.1519	13.4700
cal_consume	180.1912	200.2923	**177.6312**	201.7417
active_cal	**88.1719**	101.5719	89.2217	105.1312
walks	**522.3213**	681.3312	589.1361	703.5169
distance	**3.1429**	4.5291	3.1681	8.1796
stairs	9.4125	11.1295	**6.1956**	12.5157
active_ratio	**0.0419**	0.1115	0.0591	0.0915
nap_total_freq	**0.0396**	0.1937	0.0562	0.1449
nap_total_time	23.1660	24.9120	**21.1410**	26.1490

The results are shown in Table 2. First, the proposed Imp-GAIN apparently outperforms the original GAIN. The ratio of average improvement w.r.t Ori-GAIN is 31%. Second, our Imp-GAIN is also better than User-Avg. The average improvement ratio w.r.t. User-Avg is 6%. For those features that Imp-GAIN gets worse than User-Avg, their MAE values is close to each other (as highlighted by underline). The MAE values of Imp-GAIN-imputed features are lower than those by User-Avg. Hence, in the following analyses, we use Imp-GAIN to fill missing values of all features except for *sleep_efficiency*, *cal_consume*, *stairs*, and *nap_total_time*, which are imputed by User-Avg.

4 MODELS OF SLEEP QUALITY

We utilize the imputed Fitbit data, along with the survey data, to develop a holistic model of sleep quality over two use cases.

4.1 Sleep Efficiency Prediction

Problem Statement. For precision psychiatry, interpretability is as important as high predictive power. Given time-series data of individuals, we are interested in capturing a near future's sleep efficiency (i.e., tomorrow) based on the log from the past week. Here a standard LSTM-based model can be utilized for prediction and we add a special attention mechanism to make the model interpretable [4, 11]. LSTM is a derived model of RNN and well-known for suitable with sequential data such as time-series logs with relatively well keeping the past memories [21]. This is done in two phases as illustrated in Fig. 3: first, the LSTM-Attention block (hereafter *LA block*) and second, the ensemble of the LA blocks. We believe high predictability of sleep efficiency is able to boost the interpretability of our model, which is the main enhancement comparing to conventional regression approaches.

Phase 1. A 14-dimension multiplex vector combining 8 sleep features and 6 activity features in Table 1 was first constructed. Demographic features are utilized in Phase 2. Note that the *onbed_min* feature was excluded due to its strong correlation with other sleep features. Sequentially connected multiplex vectors $\mathbf{v}$ can be treated as one chunk c: please refer to Table 3 for the notations used in this section. Let $\mathbf{C}$ be a matrix composed of consecutive 8 $\mathbf{v}$ vectors.

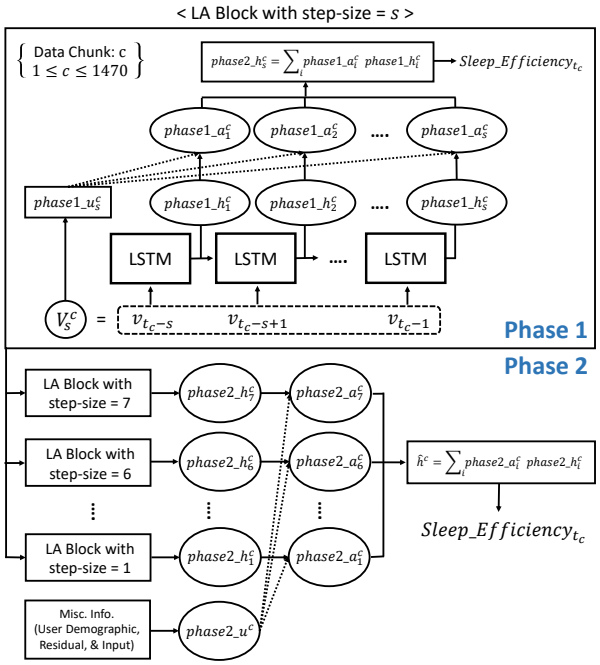


Figure 3: Our two-phase sleep efficiency prediction model.

Total 8 days are combined as one window to always include a weekend, where a window slides to the next $\mathbf{v}$ that is the next day; e.g., $\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_8}$ (denoted as $\mathbf{C_1}$) and the next day is $\mathbf{v_2}$, and considering moving one day forward is to have another data $\mathbf{C_2} = [\mathbf{v_2}, ..., \mathbf{v_9}]$. We composed the sequence of $\mathbf{C}$ per participant to represent individuals' daily log series. Hence each user will be represented by 35 sets of $\mathbf{C}$ (42-8+1=35), which becomes the batch size. The total number of input $\mathbf{C}$ is 1,470 (42users × 35$\mathbf{C}$=1470).

We next introduce LA blocks. LSTM-based models are suitable for treating sequential data. A block concept is utilized to set different step-size s per LSTM block. This allows us to concatenate different numbers of $\mathbf{v}$, starting from $\mathbf{v}_{t_c-1}$ for predicting t_c's sleep efficiency, and therefore, have different V_s^c based on s for each block, as depicted in Fig. 3. The importance of different s per block is related with our ensemble method, which will be described in Phase 2. Concatenating $\mathbf{v}$ is an important step, because it reveals which $\mathbf{v}$ (i.e., logs from a specific previous day) affects the most to sleep efficiency, with attached the attention mechanism inside a block. We set total 7 LA blocks, and each block has different step-size s ($1 \leq s \leq 7$) respectively. For instance, LA block with s=1 has only one input of $\mathbf{v}_{t_c-1}$ for LSTM and LA block with s=2 has two inputs of $\mathbf{v}_{t_c-1}$ and $\mathbf{v}_{t_c-1-1}$, respectively, and so on for the rest of s.

With respect to the attention mechanism, one literature presented the concept of a query vector ($\mathbf{u}$) to measure the similarity with the hidden vectors ($\mathbf{h}$) from one previous state [29]. We add $\mathbf{phase1\_u_s^c}$ and set as $\mathbf{phase1\_u_s^c} = tanh(V_s^c \mathbf{W_{a1}} + \mathbf{b_{a1}})$, where $\mathbf{W_{a1}}$ and $\mathbf{b_{a1}}$ are trainable parameters. Our model derives the attention score by calculating the cosine similarity $\frac{\mathbf{phase1\_u_s^c} \odot \mathbf{phase1\_h_i^c}}{||\,\mathbf{phase1\_u_s^c}\,||\;||\,\mathbf{phase1\_h_i^c}\,||}$ between query vectors $\mathbf{phase1\_u_s^c}$ and hidden vectors from LSTM

Table 3: List of notations used in predicting sleep efficiency.

Notation	Definition
$\mathbf{v}$	Daily multiplex vector composed by 8 sleep- and 6 activity-related elements
s	Step-size of a LSTM-attention (LA) block ($1 \leq s \leq 7$)
$\mathbf{C}$	Chunk matrix composed by consecutive 8 numbers of $\mathbf{v}$ (Each user has 35C)
c	Order number of C ($1 \leq c \leq 1470$ for all users)
t_c	Day of 8^{th} $\mathbf{v}$ in C with corresponding to c (t_1=8, t_{c+1}=t_c+1)
V_s^c	Concatenated set of $\mathbf{v}$ for the size of s, and updated for every c
$\mathbf{phase1\_h_i^c}$	i^{th} hidden vector of LSTM in Phase 1, and updated for every c ($1 \leq i \leq s$)
$\mathbf{phase2\_h_s^c}$	Hidden vector of LSTM weighted by the attention score in LA Block with size s
$\hat{\mathbf{h}}^{\mathbf{c}}$	Hidden vector with weighted by the attention score in Phase 2
$\mathbf{phase1\_u_s^c}$	Query vector of the attention mechanism in LA Block of size s, updated every c
$\mathbf{phase2\_u^c}$	Query vector of the attention mechanism in Phase 2, and updated for every c
$\mathbf{phase1\_a_i^c}$	Attention score vector in Phase 1, and updated for every c ($1 \leq i \leq s$)
$\mathbf{phase2\_a_i^c}$	Attention score vector in Phase 2, and updated for every c ($1 \leq i \leq 7$)
$\mathbf{W_{a1}}$, $\mathbf{W_{f1}}$	Weight matrix ($a1$: attention mechanism in Phase 1, $f1$: feed-forward network)
$\mathbf{b_{a1}}$, $\mathbf{b_{f1}}$	Bias ($a1$: attention mechanism in Phase 1, $f1$: feed-forward network in Phase 1)
se_{t_c}, $\hat{se}_{t_c}$	Given sleep efficiency on t_c and predicted sleep efficiency on t_c, respectively

$\mathbf{phase1\_h_i^c}$. The choice of cosine similarity function prevents the absolute scales from affecting results. In contrast, if we were to use dot-product functions such as $score(\mathbf{phase1\_u_s^c}, \mathbf{phase1\_h_i^c}) = \mathbf{phase1\_u_s^c} \odot \mathbf{phase1\_h_i^c}$ [25], the attention score becomes dependent on the sheer scale of $\mathbf{phase1\_h_i^c}$, irrespective of $\mathbf{phase1\_u_s^c}$ and $\mathbf{phase1\_h_i^c}$. Then $\mathbf{phase1\_h_i^c}$ can be computed as $\mathbf{phase2\_h_s^c} = \sum_{i=1}^{s} \mathbf{phase1\_a_i^c}\ \mathbf{phase1\_h_i^c}$. Lastly, $\mathbf{phase2\_h_s^c}$ is going through 1-layered feed-forward network and the predicted sleep efficiency of Phase 1 can be derived as follows: $\hat{se}_{t_c} = relu(\mathbf{phase2\_h_s^c}\ \mathbf{W_{f1}} + \mathbf{b_{f1}})$, where $\mathbf{W_{f1}}$ and $\mathbf{b_{f1}}$ are trainable parameters. Phase 1 model is trained via stochastic gradient descent (SGD) technique with dropout to minimize the loss function such that $\min_f \frac{\sum_{c=1}^{n}(se_{t_c} - \hat{se}_{t_c})^2}{n}$ meaning learning a proper objective function f towards minimizing the mean squared error (MSE) for training and testing loss.

Adding an attention mechanism allows the model to learn global trends across users. As a result, applying the learned parameters θ of an objective function f to each user's chunks will identify the most crucial previous day for predicting the target night's sleep efficiency per calculated s. In the meantime, when learning the attention score vector $\mathbf{a}$, chunks from all users are employed together, and hence, the learned $\mathbf{a}$ is universal to every user. This means although we could identify the most important previous day to predict sleep efficiency per user, the latent factors (i.e., $\mathbf{a}$) to derive that explanation may not distinctively coming from her. In order to deal with this globally-optimized issue, each user's personal traits are included as meta-data on Phase 2 to be more locally-optimized.

Phase 2. The second phase is designed to incorporate the personal traits in predicting sleep efficiency for better interpretability as well as to enhance the predictability of our model. The idea of Phase 2 is two-fold. First, in terms of personalized interpretation, the bottom part of Fig. 3 shows that a query vector $\mathbf{phase2\_u^c}$ is updated by the same equation introduced in Phase 1, with miscellaneous information (normalized to become $\in$ [0,1], with total 108 dimensions) for each user. Misc information contains demographic information, including age, BMI and ISI, and V_7^c. The dimensionality of V_7^c is 98 (7v×14features). We also have the residuals derived from Phase 1 (i,e., each seven LA blocks' minimum loss as described in Phase 1). Then, we calculate the attention scores via the same way as Phase 1 (i.e., using cosine similarity). In this way, we can expect more personalized latent factors affect the derived attention scores. Consequently, the attention scores can reveal the hidden characteristics on each user's behavioral patterns related to sleep quality.

Second, we have implemented an ensemble method to the attention mechanism. This method can improve the performance and enhance the robustness of our model, because diverse parameters θ can be learned from the same data [2, 24]. If we can make our model be more robust, it would be beneficial in alleviating the effect of noises, possibly included in the daily Fitbit logs. We can calculate $\hat{\mathbf{h}}^{\mathbf{c}}$ from $\mathbf{phase2\_a_i^c}$ as explained above and $\mathbf{phase2\_h_s^c}$, which was derived in Phase 1. After $\hat{\mathbf{h}}^{\mathbf{c}}$ going through 1-layered feed-forward network, the final sleep efficiency $\hat{se}_{t_c}$ also can be drawn by using the same equation explained in Phase 1. The model of Phase 2 is trained via SGD from the input of the hidden vectors, which are pre-trained in Phase 1. For this research we let two models learn separately (i.e., separate models) rather learn them simultaneously in an end-to-end manner. In practice, separate models can lower the model complexity so that they relax any overfitting concerns (i.e., an end-to-end model can become too complex with the scale of given data) as well as increase the learning speed.

Evaluation. For training, we utilized the first 28 sequential chunks (i.e., 80% of data) in chronological order and the remaining (20%) from the bottom for testing. If the sleep efficiency value is either (a) missing or (b) has been filled during data imputation for test data, the chunks containing those data-points were moved to the training data. Such two settings correspond to two comparing methods. When training 1,176 chunks (28 chunks×42 users) are used and when testing 294 (7×42) are used, respectively.

Predicting sleep efficiency. The major hyper-parameters for our prediction model are as follows: 1) in Phase 1, hidden layer size for LSTM; 2) in Phase 2, the number of hidden layers and their sizes for the feed-forward network to generate the query vectors; 3) in Phase 1 and 2, respectively: the size of query vectors and the hidden layer sizes for the feed-forward network that generates the prediction results. We tuned the model hyper-parameters by using grid search. In our experiments, we use AdamOptimizer with learning rate of 10^{-4}, and set dropout rate of 20%. The list of detailed parameters including parameters from seven different LA blocks are reported in the previously designated GitHub page.

As baselines, we have adopted total six existing models: 1) linear, KNN, and LASSO are used as representatives of regression methods; 2) Random Forest is used as ensemble methods; and 3) RNN and LSTM as deep-learning methods. We also prepared three different datasets based on various data imputation methods: 1) Blank: filling missing data with blank ('0') values; 2) Average: filling missing data with averaging all of previous values per feature for each user; 3) Imp-GAIN: filling missing data with using the proposed Imp-GAIN as we introduced at Section 3.

Table 4 displays the performance in Mean Absolution Error (MAE). Among existing alternatives, linear regressor showed comparatively low MSE. We conjecture this result to be due to the simple model complexity contemplating the given data size. Measuring all models including ours, our model using Phase 1 and 2 with the Imp-GAIN-imputed dataset showed the lowest test loss. In short, comparing to baselines with the averaging imputation dataset, our

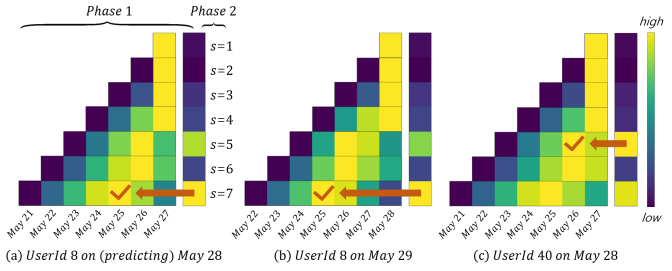
Table 4: Results in MSE by various prediction models.

Model	Blank†	Average†	Imp-GAIN†
Linear Regressor	0.01621	0.00152	0.00152
KNN Regressor	0.00464	0.00153	0.00153
LASSO Regressor	0.00521	0.00162	0.00162
Random Forest Regressor	0.02145	0.00164	0.00170
Basic RNN	0.00521	0.00162	0.00162
Basic LSTM	0.00521	0.00162	0.00162
Our Model (Phase 1)	0.00905	0.00140	0.00139
Our Model (Phase 1&2)	0.00917	0.00139	**0.00138**

† Methods of missing data imputation

model with Imp-GAIN can improve MSE around 9%-16%.[4] Moreover, we need to emphasize that while our model displayed good performance, it is also capable of interpreting the underlying causes affecting the prediction results: linear, KNN, LASSO regressors are also interpretable but they showed lower performance.

Interpreting user traits based on attention mechanism. Next, We try to interpret the attention results from our model with Phase 1 as well as Phase 2. The attention mechanism on Phase 2 is to see what step-size s would be the most explainable in predicting the target night's sleep efficiency, and the attention mechanism on Phase 1 is to see which days are more explainable within the specified s. Randomly chosen two user cases are presented in Fig. 4, and any user's trait can be explained via our attention mechanism. When observing UserId 8 to predict sleep efficiency values on May 28 and 29 (see Fig. 4(a)(b), the most explainable s were 7 on Phase 2, and the most explainable days were May 25 (i.e., the third and the fourth day, respectively from May 28). Meanwhile, UserId 40 had the different most interpretable s=5 and corresponding attention day May 26 from UserId 8 (see Fig. 4(c)).



(a) *UserId* 8 *on (predicting) May* 28 (b) *UserId* 8 *on May* 29 (c) *UserId* 40 *on May* 28

Figure 4: Step-size s and day interpretation, ranked from the attention score: test_chunk=1 (predicting May 28) for (a), (c) and test_chunk=2 (predicting May 29) for (b).

Although both attention mechanisms are originally developed to see more personal traits, they can be also used to explore general sleep patterns. Reporting ranks among s exhibit what time-frame (i.e., weekly-base) would be the most suitable for predicting sleep efficiency, in general **(Phase 2)**. For 294 test chunks (7 chunks from each 42 users), we recorded all ranks among 7 s per chunk then averaged the recorded ranks per s: the highest number of average rank among s means the corresponding s is the least important for predicting whereas the lowest number means the corresponding s is the most important. Table 5 presented the mean, SD, and 95% confidence interval of averaged rank per s (see the previously designated GitHub page to check the detailed statistics on the actual attention scores). Overall, s=7 and s=5 take the first and the second places that many users are attended on. ANOVA test ($F(6, 2051) = 36313$, p<.001) and the post-hoc Tukey's honest significant difference (HSD) test (p<.001 for all combinations) confirms the ranks of every s is statistically different from each others. Next, we investigated which days are the most crucial within the chosen s in general **(Phase 1)**, and here are the results when comparing s=3, 5, and 7 by the same way ranking s: (a) inside s=3, the first and the second ranks go to d-1 (i.e., the most recent day from the target day d) and d-2; (b) inside s=5, the first and the second ranks go to d-2 and d-3; (c) inside s=7, the first and the second ranks go to d-4 and d-3.

[4](0.00164-0.00138)/0.00164≈16%, (0.00152-0.00138)/0.00152≈9%

Table 5: Rank statistics among s using 294 test chunks.

step-size s	Mean Rank	SD	95%CI
1	6.0000	0.0000	0.0000
2	7.0000	0.0000	0.0000
3	4.9864	0.1160	0.0133
4	3.9762	0.2250	0.0258
5	1.9116	0.2844	0.0326
6	3.0374	0.1901	0.0218
7	1.0884	0.2844	0.0326

The derived attention ranks indicate two major trends. The first is related with s **(Phase 2)**. Most participants have two different periodic rhythms of sleeping, 5 days- and weekly-basis. The weekly-based pattern is a universal tendency that can be applied to contemporary life, while the 5 days-based pattern might be more distinctive traits for subjects studying in the same university and sharing similar school-related schedules. The second is related with the particular days within the certain s **(Phase 1)**. Including more previous days (i.e., increasing s) led to considering logs from further past days. This phenomenon may imply that it is natural to focus on recent days if we concern shorter period effects on shaping sleep quality: it is because sleep is affected much on the recent 1-2 days of behaviors [13]. When concerning longer period effects, periodic rhythms may start operating stronger than the effect of recent days. The current analysis may imply that when developing a physiological or psychological prediction model, it would be more effective to apply peoples' periodic behavioral rhythms such as weekly-based behavioral records than merely seeing the recent logs.

4.2 Insomnia Ranking

Problem Statement. Let ρ_i be the insomnia potential of user $i \in U$, where U is the set of users. Higher ρ_i values indicate that user i has worse sleep quality. Let $s_i \in [0, 1]$ be the sleep efficiency of user i. We define insomnia potential based on sleep efficiency: $\rho_i = 1 - s_i$. The insomnia ranking problem is defined as: given past sleep data and activity data of user set U at time $t = 1, 2, ..., T$, the goal is to generate a ranking list L for U so that those users with higher

insomnia potential at time $t = T + 1$ can be ranked at top positions in L, i.e., $L(k) \succ L(k+1)$, where $L(k)$ is the top k-th ranked user in list L, $\rho_{L(k)} > \rho_{L(k+1)}$, $k = 1, 2, ..., n-1$, and $n = |U|$.

We propose a novel ranking model, *Pairwise Learning-based Ranking Generation* (PLRG), to solve the insomnia ranking problem. The proposed PLRG model consists of four phases: (1) Ranking Pair Construction, (2) Feature Representation Learning, (3) Ranking Relation Prediction, and (4) Ranking List Generation. We elaborate these phases in the following subsection.

4.2.1 The Proposed PLRG Model. The goal is to estimate the *relative ranking relation* between users based on their feature differences. The main idea is that users with lower/higher insomnia potential tend to exhibit particular and different behaviors than those with higher/lower insomnia potential. For example, users with lower insomnia potential may exercise more (e.g., more *walks*, *distances*, and *stairs*), sleep better in past few days (e.g., better *sleep_efficiency* and more *sleep_minutes*), or sleep earlier and get up earlier in past days (e.g., lower values of *sleep_start_time* and *sleep_end_time*). Therefore, we aim to learn such kinds of sleep and activity patterns that lead to relatively lower/higher insomnia potential between users. The learned patterns, captured via feature representation, are used to predict ranking relations between users, and to generate the ranking list over all users.

Given a pair of users u and v, their *relative ranking relation* b_{uv} is defined based on insomnia potential values ρ_u and ρ_v: $b_{uv} = 1$ if $\rho_u - \rho_v \geq \tau$, and $b_{uv} = 0$ if $\rho_u - \rho_v < \tau$, where $b_{uv} = 1$ indicates u has higher insomnia potential than v, and τ is the threshold that determines the relative ranking relation according to the difference between insomnia potential values. Higher τ values refer to strict ranking relations. Note that since the sleep efficiency values are close, we need to set the threshold τ to be a small value (e.g., $\tau = 0.006$ by default). We will present how τ affects the performance.

Phase 1: Ranking Pair Construction. Given the features from sleep and activity data in past T days, we construct data instances to train a prediction model of ranking relation. A user pair is used to construct a data instance. Let $\mathbf{x}^u$ and $\mathbf{x}^v$ be the feature vectors of users u and v. A data instance, denoted by $\mathbf{x}^{uv}$, is made up of a vector of their element-wise feature differences $\mathbf{x}^u - \mathbf{x}^v$ and their corresponding ranking relation b_{uv}, i.e., $\mathbf{x}^{uv} = \langle \mathbf{x}^u - \mathbf{x}^v, b_{uv} \rangle$. We use $\mathbf{x}^{uv}$ of all pairs of users $u, v \in U$ in past days for training.

Phase 2: Feature Representation Learning. The second phase is to learn *feature representation* of each user pair from the raw feature vector $\mathbf{x}^{uv}$ to capture the sleep and activity patterns betwen users. Let $\mathbf{x}_i^{uv}$ (abbreviated as $\mathbf{x}_i$ thereafter) be the feature vector of users u and v in the past i-th day ($i = 1, 2, ..., T$). We aim to make the feature representation capture the two user behaviors. The first is the day-by-day *sequential* change of pairwise behaviors from the first to the latest observed days. The second is the *interaction* effect between sleep efficiency and other features since sleep efficiency is directly relevant to the ranking of insomnia potential. To realize such idea, we *pre-train* a model that can generate the feature representation of each user pair from $\mathbf{x}_i$.

The overview of the proposed pre-trained model for feature representation learning is shown in which the black and grey colored parts of Fig. 5. The right-bottom part is to learn the sequential features while the left-bottom part is to learn the feature interactions.

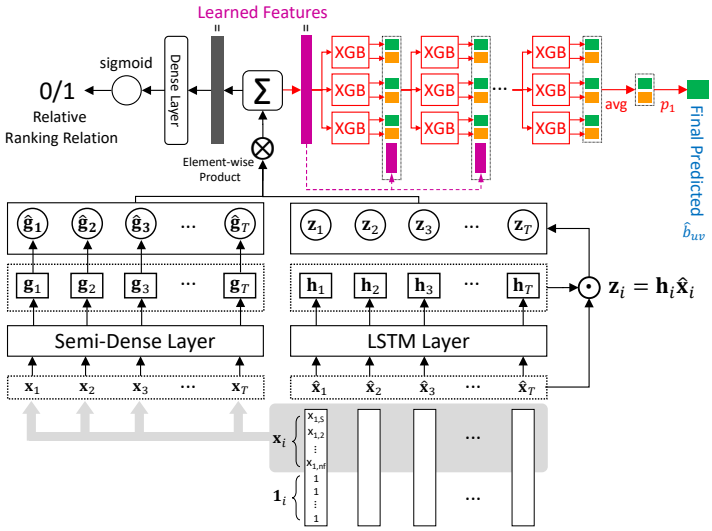


Figure 5: Flowchart of Phase 2 and Phase 3 in the proposed PLRG model. Phase 2 are colored by black and gray (bottom-left, bottom-right, and top-left), and Phase 3 are highlighted by rich colors (top-right).

We combine the outputs of both parts, along with a dense layer and the sigmoid activation function (top-left) to pre-train and learn the prediction of relative ranking relations.

First, we learn the sequential features by each day's feature vectors $\mathbf{x}_i$ as the input. Since the sequential features will be combined with the interaction features (bottom-left), we concatenate each $\mathbf{x}_i$ with a vector $\mathbf{1}_i$ with $n_f - 1$ 1s, where n_f is the number of features. The new vector is denoted as $\hat{\mathbf{x}}_i = [\mathbf{x}_i, \mathbf{1}_i]$, whose dimensionality is $2n_f - 1$. Then we leverage *Long Short-Term Memory* (LSTM) architecture [21] to learn sequential features by feeding $\hat{\mathbf{x}}_i$ as the input. With LSTM modeling, we collect the output of each LSTM hidden state: $\mathbf{h}_1 = g(\mathbf{x}_1, \mathbf{h}_0, \mathbf{c}_0, \Theta)$ and $\mathbf{h}_i = g(\mathbf{x}_i, \mathbf{h}_{i-1}, \mathbf{c}_{i-1}, \Theta)$, $i = 2, 3, ..., T$, where $\mathbf{x}_i$, $\mathbf{h}_i$, and $\mathbf{c}_i$ are the input, output, and cell vectors at time i, respectively, and Θ is the set of parameters of LSTM neurons. Here we use *sigmoid* as the activation function. During the this process, the initial values of the hidden states $\mathbf{h}_0$ and the memory cells $\mathbf{c}_0$ are set to zero. We use the hidden state's output vector $\mathbf{h}_i$ to weight the input vector $\hat{\mathbf{x}}_i$. The sequential features denoted by $\mathbf{z}_i$ can be generated via $\mathbf{z}_i = \mathbf{h}_i \hat{\mathbf{x}}_i$.

The second part is to learn the interactions between sleep efficiency and other features as *interaction features*. Let $x_{i,j}$ be feature j's value at time i, where $j = 1, 2, ..., n_f$ and $j = 1$ (also written as $j = S$ for clearness) is referred as sleep efficiency. The input feature vector $\mathbf{x}_i$ is fed into a newly defined *semi-dense layer*, whose output vector is denoted by $\mathbf{g}_i$. Vector $\mathbf{g}_i$ is further concatenated with n_f 1s and derive new vector $\hat{\mathbf{g}}_i$, in order to combine with $\mathbf{z}_i$. The semi-dense layer accepts $\mathbf{x}_i$ and generates $\mathbf{g}_i$ via: $g_{i,j} = w^i_{S,j} x_{i,S} + w^i_{j,j} x_{i,j}$, where $j = 2, 3, ..., n_f$, $w^i_{S,j}$ is the contribution of sleep efficiency $x_{i,S}$ to feature $x_{i,j}$ at time i, and $w^i_{j,j}$ is the contribution of feature $x_{i,j}$. In other words, this equation is devised to learn the weights depicting interactions between sleep

efficiency and other features. Eventually we can obtain $\hat{\mathbf{g}}_i$, whose dimensionality is $2n_f - 1$.

We use element-wise product $\otimes$ to combine sequential feature vector $\mathbf{z}_i$ with interaction feature vector $\hat{\mathbf{g}}_i$. The final learned vector of feature representation $\mathbf{f}$ can be derived by summing up each feature over time $i = 1, 2, ..., T$. Such two actions can be described by: $\mathbf{f} = \sum_{i=1}^{T} \mathbf{z_i} \otimes \hat{\mathbf{g}}_i$. The feature vector $\mathbf{f}$ is fed into a dense layer, along with a sigmoid activation function, and expects that the ranking relation b_{uv} can be predicted. To derive the feature representation $\mathbf{f}$, we use training data to learn all model parameters $\{\Theta, \mathbf{W}, \Phi\}$, where $\mathbf{W} = [w^i_{S,j}, w^i_{j,j}]$, and Φ is the set of parameters before and after the dense layer. Binary cross entropy and Adam optimizer with learning rate of 10^{-3} are used for pre-training.

Phase 3: Ranking Relation Prediction. With the derived feature vector $\mathbf{f}$, we predict the final relative ranking relation by combining two the state-of-the-art deep classification models, *Deep Forest* [31] and *XGBoost* [9]. The main reason we resort to Deep Forest, rather than simply using Phase 2, is two-fold. First, we do not have a large-scale dataset that is needed to train a good deep neural network. Second, a neural network may have too many hyper-parameters to tune. Deep forest with the cascade structure adaptively goes deep through cross validation, and thus possesses the representation learning ability and reduce the risk of overfitting. In addition, we choose XGBoost as the kernel of deep forest since it has lower model variance than other tree-based methods (see top-right of Fig. 5). We choose three-fold cross validation, three XGBoost per layer, and use default parameters in XGBoost. We do not use the predict binary labels 0 or 1 as the predicted $\hat{b}_{uv}$. Instead we use the generated probability $\hat{b}_{uv} = p_1$, where p_1 is the probability of prediction as "1". That said, the predicted $\hat{b}_{uv}$ indicates the probability that u is ranked higher than v.

Phase 4: Ranking List Generation. The task aims at generating the final ranking list of users using the predicted $\hat{b}_{uv}$ of all user pairs (u, v). The method has two steps. First, we construct a bi-directed complete graph $G = (U, E)$, in which each node $u \in U$ is a user, and each directed edge $e = (u, v) \in E$ is associated with a weight value $\omega_{uv} = 1 - \hat{b}_{uv}$. Lower weights ω_{uv} mean higher possibility that user u is ranked higher than user v. Second, we consider the ranking list generation problem as finding a path $\mathcal{P}$ with length $|U| - 1$ that is required to cover all nodes in G, i.e., $\forall u, v \in \mathcal{P} : u \neq v$. That said, we model the ranking list generation as the *Asymmetric Traveling Salesman Problem* (ATSP) [18]. A cost-based greedy heuristic algorithm [18] can be used to solve ATSP and find the directed path as the final ranking list $\hat{L}$.

4.2.2 Evaluation. We conduct experiments to evaluate the proposed PLRG. Six compared methods are considered. (1) RankNet [6] and (2) LambdaMART [6]: the state-of-the-art learning-to-rank methods, (3) LSTM-R: feeding *each user u's feature vector* $\mathbf{x}^u$ into LSTM to predict the *real values* of insomnia potential that is further used for user ranking, (4) LSTM-B: feeding *user-pair feature feature vectors* $\mathbf{x}^{uv}$ into LSTM to predict ranking relations used for our Phase 4's ranking (i.e., without using Deep Forest with XGBoost), (5) DF-XGB: feeding *raw user-pair feature vectors* $\mathbf{x}^{uv}$ into Deep Forest (XGBoost as the kernel) to predict ranking relations used for our Phase 4's (i.e., only Phase 3 & 4 without learning sequential features and interaction features). (6) The proposed PLRG.

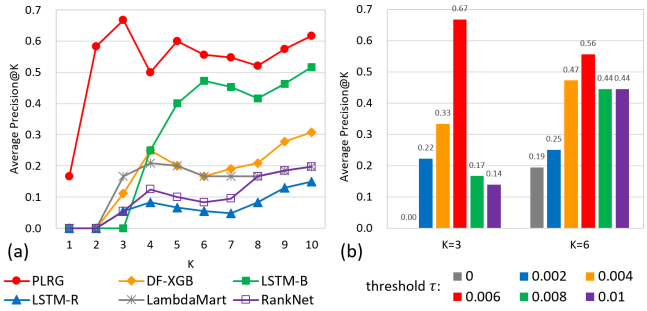


Figure 6: Experimental results of ranking: (a) Precision by varying K for different methods (with fixed $\tau = 0.006$). (b) Precision@K by varying threshold τ with $K = 3$ and $K = 6$.

In the evaluation settings, a data instance contains a feature vector with past 8 consecutive days ($T = 8$) of a user, and the next following day's (i.e., 9-th day) sleep efficiency is used to obtain the insomnia potential ρ and the ground-truth ranking list L. Time window was moved day-by-day to generate all data instances. The first 80% sleep and activity data instances were used for training and the remaining 20% were used for testing. We report the *Average Precision@K* (*AP@K*), as the evaluation metric, i.e., $AP@K = \frac{|\hat{L}^K \cap L^K|}{K}/N$, where $K = 1, 2, ..., 10$, $\hat{L}^K$ and L^K are the sets of predicted and ground-truth top-K users, respectively, and N is the number of testing data instances.

Experimental Results. We draw several insights from Fig. 6. First, we can find that our PLRG leads to the highest precision scores, especially when $K \leq 5$. Such results demonstrate that PLRG is able to more accurately detect users with higher insomnia than competing methods. Second, the compared methods cannot generate accurate ranking when K is small, but their precision scores get better when $K > 5$. Third, the next two models with higher precision scores are LSTM-B and DF-XGB, which corresponds to two main components (i.e., feature representation learning, and Deep Forest with XGBoost) in PLRG. Comparing LSTM-B and DF-XGB with PLRG implies the predictability can get boosted once (a) the features can be better learned and (b) a proper prediction model is adapted to avoid overfitting. Fourth, directly predicting the real values of insomnia potential via LSTM (i.e., LSTM-R) is worse. It may be because the differences between insomnia values are very small. Values being accurately inferred (i.e., Section 4.1) do not imply the insomnia risk ranking of users can be precisely generated. Last, the state-of-the-art learning-to-rank models, LambdaMART and RankNet, also fail to predict the ranking. Compared with LambdaMART, our PLRG leads to around 300% improvement in *AP*@3 and around 233% improvement in *AP*@6. In Fig. 6(b), we report the scores of *AP*@3 and *AP*@6 of our PLRG by varying the threshold τ that determines the relative ranking relation b_{uv}. Note that $\tau = 0.006$ leads to the highest precision. Smaller τ makes the ranking relation hard to distinguish users from each other, and larger τ generates the imbalance of data instances between $b_{uv} = 1$ and $b_{uv} = 0$.

5 CONCLUSION AND DISCUSSION

Mental well-being is fundamental to human health. Heterogeneous data sources that are widely becoming available can make a huge impact in psychiatry. This study, base on real-world logs gathered from insomnia sufferers over a 6-week period, demonstrates how sleep and activity data collected from smart bands can be analyzed to estimate sleep quality (that have long been measured via self-reported questionnaires). In this process, we notice that missing data handling becomes a key challenge and propose to impute data via an improved generative adversarial networks, called Imp-GAIN. Then we present two specific models that give the rich context of sleep and activity relationships. Our LSTM-Attention ensemble model showed the best performance in predicting the next night's sleep efficiency and also was capable of interpreting the underlying grounds for results. Our pairwise learning-based ranking generation model, called PLRG, could rank individuals who will be at high risk of insomnia in the next sleep, based on representation learning of sequential and interaction features. These models and prediction outcomes have been reviewed by a psychiatrist for a practical use and plan to be used at a hospital for a trial.

This work has several limitations that could be addressed in the future. First, participants were recruited from the same university and hence may induce sampling biases. Note that this homogeneity, however, benefited our study from controlling various exogenous variables such as weather, holidays, school events, and so on. Second, a larger-scale study will help us draw conclusions that are applicable for the general public. Future studies may target generic patients who visit hospitals regularly, to ensure the generality of our prediction models. Third, the gathered daily logs were grouped in the form of a chunk – 8 day period – in this research to always include a weekend. This size was appropriate for the 6-week log. We may in the future utilize even longer chunk sizes, if we were to conduct a longitudinal study.

ACKNOWLEDGMENTS

We thank Sungwon Park for his contribution on evaluating baseline models. This research was supported by Basic Science Research Program (No. NRF-2017R1E1A1A01076400) and Next-Generation Information Computing Development Program (No. NRF-2017M3C4A7063570) through the National Research Foundation of Korea funded by the Ministry of Science and ICT in Korea. This work was also supported by Ministry of Science and Technology (MOST) Taiwan with grants 108-2636-E-006-002 (MOST Young Scholar Fellowship Program) and 107-2218-E-006-040, and supported by Academia Sinica Thematic Research Program with grant AS-107-TP-M05.

REFERENCES

[1] Rajendra Acharya, Oliver Faust, N Kannathal, TjiLeng Chua, and Swamy Laxminarayan. 2005. Non-linear analysis of EEG signals at various sleep stages. *Elsevier Computer Methods and Programs in Biomedicine* 80, 1 (2005), 37–45.

[2] Bo An, Haipeng Chen, Noseong Park, and VS Subrahmanian. 2016. MAP: Frequency-based maximization of airline profits based on an ensemble forecasting approach. In *proc. of the ACM SIGKDD.*

[3] Jutta Backhaus, Klaus Junghanns, Andreas Broocks, Dieter Riemann, and Fritz Hohagen. 2002. Test–retest reliability and validity of the Pittsburgh Sleep Quality Index in primary insomnia. *Journal of Psychosomatic Research* 53, 3 (2002).

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[5] Tessa F Blanken, Jeroen S Benjamins, Denny Borsboom, Jeroen K Vermunt, Casey Paquola, Jennifer Ramautar, Kim Dekker, Diederick Stoffers, Rick Wassing, Yishul Wei, et al. 2019. Insomnia disorder subtypes derived from life history and traits of affect and personality. *Elsevier Lancet Psychiatry* (2019).

[6] Chris J.C. Burges. 2010. From RankNet to LambdaRank to LambdaMART: An Overview. *Technical Report, Microsoft Research* (2010).

[7] Daniel J Buysse, Charles F Reynolds III, Timothy H Monk, Susan R Berman, and David J Kupfer. 1989. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry Research* 28, 2 (1989), 193–213.

[8] Danilo Bzdok and Andreas Meyer-Lindenberg. 2017. Machine learning for precision psychiatry: Opportunites and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* (2017).

[9] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *proc. of the ACM SIGKDD.*

[10] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer representation learning for medical concepts. In *proc. of the ACM SIGKDD.*

[11] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems.* 3504–3512.

[12] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. 2008. Querying and mining of time series data: experimental comparison of representations and distance measures. In *proc. of the VLDB.*

[13] David F Dinges, Frances Pack, Katherine Williams, Kelly A Gillen, John W Powell, Geoffrey E Ott, Caitlin Aptowicz, and Allan I Pack. 1997. Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4–5 hours per night. *Sleep* 20, 4 (1997), 267–277.

[14] Kelly R Evenson, Michelle M Goto, and Robert D Furberg. 2015. Systematic review of the validity and reliability of consumer-wearable activity trackers. *IJBNPA* 12, 1 (2015), 159.

[15] Brisa S Fernandes, Leanne M Williams, Johann Steiner, Marion Leboyer, André F Carvalho, and Michael Berk. 2017. The new field of 'precision psychiatry'. *BMC medicine* 15, 1 (2017), 80.

[16] Nancy L Galambos, Andrea L Dalton, and Jennifer L Maggs. 2009. Losing sleep over it: Daily variation in sleep quantity and quality in Canadian students' first semester of university. *Journal of research on adolescence* 19, 4 (2009), 741–761.

[17] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *proc. of the NIPS.*

[18] Gregory Gutin and Abraham P. Punnen. 2007. The Traveling Salesman Problem and Its Variations. *Combinatorial Optimization* (2007).

[19] J Heikenfeld, A Jajack, J Rogers, P Gutruf, L Tian, Tingrui Pan, R Li, M Khine, J Kim, and J Wang. 2018. Wearable sensors: modalities, challenges, and prospects. *Lab on a Chip* 18, 2 (2018), 217–248.

[20] Wahyu Hidayat, Toufan D Tambunan, and Reza Budiawan. 2018. Empowering Wearable Sensor Generated Data to Predict Changes in Individual's Sleep Quality. In *proc. of the IEEE ICoICT.*

[21] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[22] David A. Kalmbach, J. Todd Arnedt, Vivek Pillai, and Christopher L. Drake. 2016. Identifying At-Risk Individuals for Insomnia Using the Ford Insomnia Response to Stress Test. *Sleep* 39, 2 (2016), 449–456.

[23] Uri Kartoun, Rahul Aggarwal, Andrew Beam, Jennifer K. Pai, Arnaub Chatterjee, Timothy P. Fitzgerald, Isaac Kohane, and Stanley Shaw. 2018. Development of an Algorithm to Identify Patients with Physician-Documented Insomnia. *Scientific Reports* 8 (2018), Article Number: 7862.

[24] Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *proc. of the IEEE CVPR.*

[25] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[26] Hannah Morphy, Kate M Dunn, Martyn Lewis, Helen F Boardman, and Peter R Croft. 2007. Epidemiology of insomnia: a longitudinal study in a UK population. *Sleep* 30, 3 (2007), 274–280.

[27] Sarun Paisarnsrisomsuk, Michael Sokolovsky, Francisco Guerrero, Carolina Ruiz, and Sergio A. Alvarez. 2018. Deep Sleep: Convolutional Neural Networks for Predictive Modeling of Human Sleep Time-Signals. In *KDD Deep Learning Day.*

[28] Aarti Sathyanarayana, Jaideep Srivastava, and Luis Fernandez-Luque. 2017. The science of sweet dreams: predicting sleep efficiency from wearable device data. *Computer* 50, 3 (2017), 30–38.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems.* 5998–6008.

[30] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GAIN: Missing Data Imputation using Generative Adversarial Nets. In *proc. of the ICML.*

[31] Zhi-Hua Zhou and Ji Feng. 2017. Deep Forest: Towards An Alternative to Deep Neural Networks. In *proc. of the IJCAI.*