A Two-stage Imbalanced Learning Method for Sleep Stages Classification using Consumer Activity Trackers

Zilu Liang[1,2][a] and Mario Alberto Chapa-Martell[3][b]
[1]*School of Engineering, Kyoto University of Advanced Science, Kyoto, Japan*
[2]*Institute of Industrial Science, The University of Tokyo, Tokyo, Japan*
[3]*Silver Egg Technology, Osaka, Japan*
z.liang@cnl.t.u-tokyo.ac.jp, mchapam0300@gmail.com

Keywords: Fitbit, Activity Tracker, Sleep Stage, Machine Learning, Imbalanced Learning.

Abstract: Consumer sleep tracking technologies such as Fitbit activity trackers are increasingly used in scientific studies to measure sleep, but these devices are known to be inaccurate for measuring sleep stages. In this study we propose a two-stage imbalanced learning method to improving Fitbit accuracy. The stage-1 model classifies a Fitbit measurement into either correct or incorrect. If the measurement is classified as incorrect, then the stage-2 model corrects it by re-classifying it into one of the four sleep stages. We reliably examined the performance of different combinations of machine learning techniques (i.e. Naive Bayes, random forest and support vector machine) and resampling techniques (i.e. up sampling and down sampling) through leave-one-out nested cross validation. The results showed that using Naive Bayes as the machine learning technique in both stages achieved the best performance, and down sampling needed to be applied during the training of stage-1 model. Our proposed model successfully improved Cohen's Kappa by up to 27% and Matthews correlation coefficient (MCC) by up to 26%. Performance improvement was achieved mainly through improving the accuracy for light sleep (by 29%). The proposed method can be used to post-process data from Fitbit activity trackers to achieve better accuracy in sleep staging.

1 INTRODUCTION

The recent decade has witnessed the rise of consumer sleep tracking technologies such as Fitbit activity trackers (Liu, Ploderer, & Hoang, 2015). These devices are increasingly used in scientific studies to measure sleep patterns because they are affordable, non-invasive and are easy to use for longitudinal collection of sleep data in natural environments (Bian, Guo, Xie, Parish et al., 2017; Weatherall, Paprocki, Meyer, Kudel, & Witt, 2018; Weaver et al., 2018). Despite of these attractive features, Fitbit devices have raised wide concern regarding data quality, as many validations studies revealed that Fitbit devices are inaccurate compared to clinical sleep monitors especially for measuring sleep stages (M De Zambotti, Cellini, Goldstone, Colrain, & Baker, 2019; Liang & Chapa-Martell, 2018b, 2019a; Liang & Ploderer, 2017). Previous models of Fitbit rely solely on arm movement patterns measured by embedded accelerometer to infer sleep and wake, and they tend to overestimate sleep and underestimate wakefulness (Meltzer, Hiruma, Avis, et al., 2015). The latest models such as Charge 2 and Charge 3 take advantage of multiple streams of signals including arm movement and heart rate to detect four sleep stages (i.e. wakefulness, light sleep, deep sleep and REM sleep). Nevertheless, these devices only achieved an accuracy of 60% for light sleep, deep sleep and REM sleep (Massimiliano De Zambotti, Goldstone, Claudatos, et al., 2017), and they demonstrated mediocre accuracy on wakefulness (30%-67%) (Massimiliano De Zambotti et al., 2017; Liang & Chapa-Martell, 2018b, 2019c). Since Fitbit sleep data may not accurately reflect the real sleep patterns being measured, these data can mislead researchers and compromise the reliability of their research outcomes. To this end, enhancing the accuracy of Fitbit for measuring sleep stages can benefit researchers who intend to use these devices in scientific studies as well as individual users who rely on Fitbit data to make self-care decisions.

[a] https://orcid.org/0000-0002-2328-5016
[b] https://orcid.org/0000-0002-4110-4346

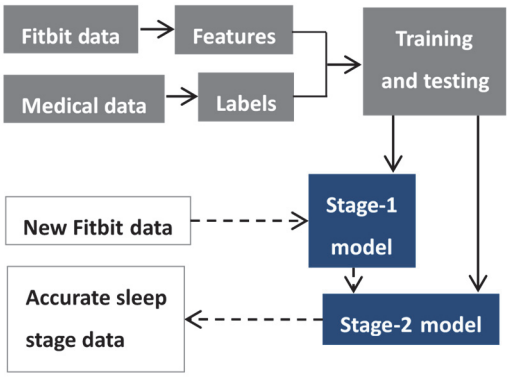

Figure 1: Outline of the proposed two-stage model.

The objective of this study was to develop a computational method that post-processes Fitbit data to achieve better accuracy for sleep stages. As illustrated in Figure 1, our method consists of two stages of classifications. The stage-1 model classifies Fitbit sleep data into "correct" (negative class) or "incorrect" (positive class). The stage-2 model re-classifies incorrect instances into one of the four classes: "deep sleep" (class 1), "light sleep" (class 2), "REM sleep" (class 3) or "wakefulness" (class 4). We examined the performance of different combinations of supervised machine learning techniques, including Naïve Bayes (NB), random forest (RF) and support vector machine (SVM). One major challenge in supervised learning is the collection of ground truth/labels. In this study, we leveraged epoch-wise sleep stage data that was simultaneously measured with a medical-grade sleep monitor as the ground truth. Due to class imbalance, we also applied random up sampling or random down sampling techniques to even the class frequency. These techniques were selected as they have shown to outperform other techniques for sleep stage classification (Liang & Chapa-Martell, 2019b; Xiao, Yan, Song, Yang, & Yang, 2013; Zhu, Li, & Wen, 2014).

The contribution of this paper is two-fold. First, we introduced a two-stage classification model that post-processes Fitbit data to achieve more accurate classification for sleep stages. Second, we reliably evaluated model performance using multiple measures through nested cross validation. The models were tested on a whole night of sleep data rather than a set of random sleep epochs. The results thus reflect model performance in real situations. The rest of the paper is organized as follows. In Section 2 we present related work on sleep stage classification in clinical settings and the validation of Fitbit sleep data. In Section 3 we describe the proposed method in detail. Performance evaluation is presented in Section 4. In Section 5 we discuss the interpretation of the results and point out directions for future research. The paper is concluded in Section 6.

2 RELATED WORK

The history of the gold standard sleep test—polysomnography (PSG)—can be traced back to 1960s. A PSG test involves monitoring physiological changes during sleep using surface electrodes including electroencephalography (EEG), electrooculography (EOG) and electromyography (EMG). These are the main signals used to infer sleep stages. Other signals, such as nasal airflow, abdominal effort, blood oxygenation, are mainly used to diagnose sleep disorders such as sleep apnoea.

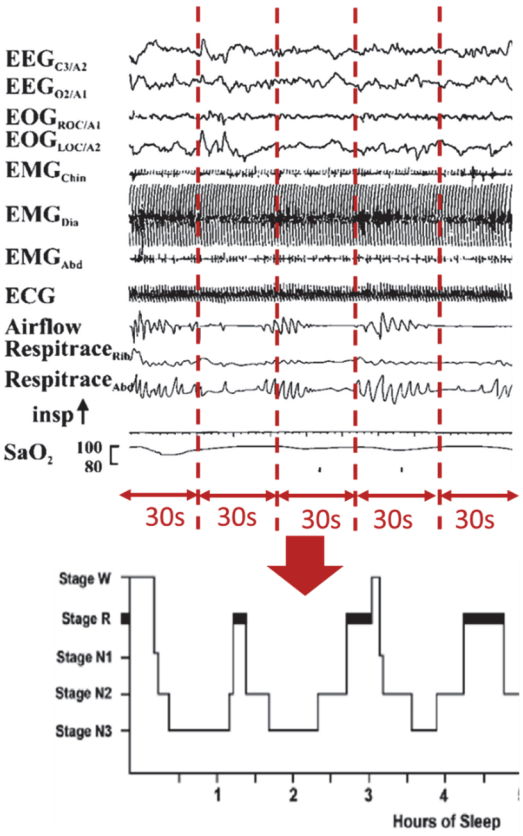

Figure 2: Sleep staging in clinical settings.

The process of identifying sleep stages from multiple streams of physiological signals is called sleep scoring or sleep staging, which is a critical step in PSG analysis. Sleep staging is usually conducted by a human expert based on epoch-by-epoch visual interpretation of the signals following established guidelines (Ancoli-Israel, Chesson, & Quan, 2017).

As illustrated in Figure 2, the signals are firstly divided into 30-second epochs. Commencing at the start of the recording, the epochs are annotated sequentially using 5 labels: wake, stage1, stage2. Stage3, stage4. If two or more stages coexists during a single epoch depends on the signal features, the stage comprising the dominant portion is assigned. The final output of sleep staging is a hypnogram—a plot of sleep stages as a function of time. The process of manual sleep staging is complex and time-consuming, which limits the capacity of many hospitals in offering such examination service (Colten & Altevogt, 2006). The disadvantage associated with PSG brings the need to develop alternative sleep-tracking technologies that are easier to use, more affordable and more accessible. Portable electroencephalogram (EEG) and actigraphy are some of the alternatives that are more comfortable to wear and can be used in natural settings to collect longitudinal data. That said, these devices are still expensive and are limited to clinical use only.

Personal sleep tracking did not become possible until off-the-shelf wearable trackers entered the consumer market. Fitbit is one of the major manufacturers of wearable trackers. Previous studies have intensively investigated the accuracy (M De Zambotti et al., 2019; Massimiliano De Zambotti et al., 2017; Liang & Chapa-Martell, 2018b, 2019a), usability (Liang, Ploderer, et al., 2016; Liang & Ploderer, 2017), and health impact of these sleep trackers (Liang, Chapa-Martell, & Nishimura, 2016a, 2016b; Liang & Ploderer, 2017; Shelgikar, Anderson, & Stephens, 2016). In recent years, Fitbit devices are increasingly used in scientific studies to measure sleep outcomes (Bian et al., 2017; Weatherall et al., 2018; Weaver et al., 2018). Compared to clinical actigraphy, Fitbit's advantage of combining more than one source of information is also recognized by the medical community (Goldstone, Baker, & De Zambotti, 2018). Nevertheless, the accuracy of Fitbit devices is not satisfactory, especially the accuracy for measuring sleep stages. Previous validation studies found that the typical accuracy of Fitbit for measuring wakefulness was between 30-40%, and the accuracy for measuring light sleep, deep sleep and REM sleep was approximately 60% (M De Zambotti et al., 2019; Liang & Chapa-Martell, 2018b). A number of research projects have attempted to develop new sleep tracking devices (Nam, Kim, & Lee, 2016; Rahman et al., 2015). Given the large user base of Fitbit, however, a more practical approach is to leverage the data measurable by Fitbit devices and to propose new algorithms that post-process Fitbit data for more accurate sleep staging. Backed by this rationale, we proposed a two-stage model to achieve the goal. Our method significantly deviates from the mainstream and we present promising results in this paper.

3 METHOD

3.1 Data Preparation

To prepare training dataset, we used a Fitbit Charge 2 and a medical-grade single channel EEG concurrently to measure a whole night of sleep from 23 participants. We also collected demographic information such as age, gender, and subjective sleep quality using the Pittsburgh Sleep Quality Index (PSQI) questionnaire (Buysse, Reynolds, Monk et al., 1989).

Fitbit provides two types of data. One type is daily aggregated data including minutes asleep, minutes awake, and sleep efficiency. These data can be retrieved through Fitbit Public API (application programming interface). Another type is intraday data including time series of sleep stages and heart rate. These data can be retrieved through Fitbit Partner API at high resolution (below 10 seconds). A C# program was developed to synchronize all streams of data and to interpolate the data at a resolution of 1 second. The data were then aggregated to 30-second epochs following clinical standard, and each epoch was mapped to one instance in the final dataset.

The medical data was sent back to the company for analysis. The raw signals were firstly scored automatically by proprietary software that complies with clinical sleep scoring standard (Ancoli-Israel et al., 2017). The results were then inspected by sleep experts and revised if necessary. The final data consist of 23 nights of sleep hypnograms at 30-second resolution. The medical data was then synchronized with the processed Fitbit intra-day sleep data and heart rate data.

3.2 Model Construction

3.2.1 Overview

The problem of interest was treated as a two-stage classification problem. Stage-1 classification is a binary classification task that aims to predict if a Fitbit measurement is correct (negative class) or not (positive class) compared to the medical data. If the measurement is incorrect, stage-2 classification will correct it by reclassifying it into one of the four sleep stages (1=deep sleep, 2=light sleep, 3=REM sleep,

4=wakefulness). The denotations used throughout this paper are summarized in Table 2.

3.2.2 Feature Extraction and Labelling

Table 2: Denotations used in this paper.

Denotation	Description
n	Sleep epoch ID.
x_n^{FS}	Original Fitbit sleep stage classification for epoch n.
x_n^{HR}	Fitbit average heart rate data for epoch n.
Δx_n^{HR}	Change in heart rate in epoch n compared to previous epoch (= $x_n^{HR} - x_{n-1}^{HR}$).
W	Wakefulness.
L	Light sleep.
D	Deep sleep.
R	REM sleep.
ACC_k	Accuracy for sleep stage k, $k \in \{W, L, D, R\}$.
PRE_k	Precision for sleep stage k, $k \in \{W, L, D, R\}$.
Kappa	Cohen's Kappa.
MCC	Generalization of Matthews correlation coefficient to multiclass cases.
$\boldsymbol{M}_{ij}$	Overall performance of a combination of machine learning techniques i and j for stage-1 and stage-2 respectively (= $\{Kappa_{ij}, MCC_{ij}\}$) .

We created a set of 21 features using the raw data from Fitbit and PSQI questionnaire, which can be easily obtained in both clinical and daily life settings. The whole set of features was used in both Stage-1 and Stage-2 classification. We first segmented one night of sleep data and heart rate data into 30s epochs. Time-dependant features were then calculated for each epoch. These features included sleep stage measured by Fitbit and averaged over the n-th epoch x_n^{FS}, heart rate measured by Fitbit and averaged over the n-th epoch x_n^{HR}, sleep stages averaged in previous three epochs x_{n-1}^{FS}, x_{n-2}^{FS}, x_{n-3}^{FS}, sleep stages averaged in three subsequent epochs x_{n+1}^{FS} , x_{n+2}^{FS} , x_{n+3}^{FS} , changes in heart rate Δx_n^{HR} andΔx_{n+1}^{HR}, and epoch ID n. Time-independent features are then merged with time variant features. Three time-independent features, including sex, age, and PSQI score, can be obtained using the PSQI questionnaire. Other time-independent features such as total sleep time, total wake time, sleep efficiency, and ratio of each sleep stage can be obtained from daily aggregated Fitbit sleep data. These demographic information and sleep pattern information were used as features because they were shown to affect device accuracy in previous validation studies (Liang & Chapa-Martell, 2018a, 2019a).

The labels were obtained using the following conversion method. For stage-1 binary classification, the label for each instance was obtained by comparing the Fitbit sleep stage data and the medical data corresponding to that epoch. If Fitbit sleep stage equals medical sleep stage, the label was 0; otherwise, the label was 1. For stage-2 multiclass classification, the medical data was used as the labels.

3.2.3 Machine Learning and Resampling Techniques

Three machine learning algorithms were applied: Naive Bayes (NB), random forest (RF), and support vector machine (SVM) with linear kernel. These techniques have demonstrated good performance in sleep stage classification in clinical settings (Liang & Chapa-Martell, 2019b; Xiao et al., 2013; Zhu et al., 2014). In each stage of the model, one of the three machine learning algorithms was used to perform the classification task.

To address the imbalanced nature of the dataset, we also applied resampling techniques—random up sampling and random down sampling—to balance the frequency of each class. Our previous pilot study suggested that these two simple resampling techniques demonstrated better performance than more complicated resampling techniques such as SMOTE and ROSE (Liang & Chapa-Martell, 2019b, 2019d). It is worth noting that resampling was performed inside cross validation to ensure that the test results are not biased.

3.2.4 Model Tuning and Testing

Stage-1 and stage-2 models were trained using the same training set but with different labels, as the former is a binary classification task and the latter is a four-category classification task. The 21 features were extracted and paired up with the corresponding labels. All features were normalized over the whole training set. The optimal values for model parameters were decided using grid search with 10-cross fold validation (repeated three times). For Naïve Bayes models, Laplace smoother was set between (0, 5) in increments of 0.5, and the bandwidth of the kernel density was set between (0, 5) in increments of 1. For random forest, the number of predictors sampled for splitting at each node was set between [1, 20] in increments of 1. For support vector machine, penalty parameter C was selected from the set (0.001, 0.01, 0.1, 1, 10, 100) and linear kernel was used.

We used the leave-one-out strategy for model testing. This was to ensure that the test set resembled the class distribution in real situations and the models were tested on a whole night of sleep rather than a random set of sleep epochs. Since sleep data was collected from 23 participants, model training and testing were repeated 23 times with one participant's data being kept apart each time as test set. The data of the rest 22 participants were merged into a large training set. Models were trained on the training set and parameters were tuned through 10-fold cross validation. Resampling was performed on the training set inside cross validation and was not performed on the test set to avoid performance bias. Performance measures were averaged over the 23 repetitions to generate final performance. Figure 3 illustrates the leave-one-out nested cross validation.

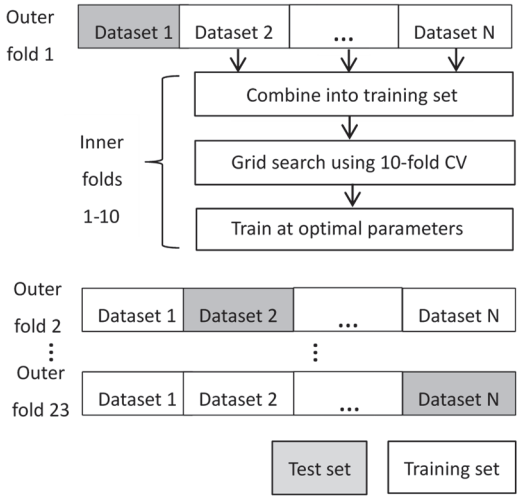

Figure 3: Leave-one-out nested cross-validation.

3.3 Performance Measures

We used three measures suited for imbalanced and multiclass problems to thoroughly evaluate the performance of different combinations of machine learning and resampling techniques. Cohen's *Kappa* statistic was used to evaluate how much better a classifier performed over random guess according to the frequency of each class (J. Landis & G. Koch, 1977). A generalization of Matthews correlation coefficient (*MCC*) to multiclass cases (Gorodkin, 2004) was used to evaluate the overall performance of models by considering true and false positives and negatives. We also examined the accuracy (ACC_k) and precision (PRE_k) of models for detecting each sleep stage, $k \in \{W, L, D, R\}$ where W, L, D, R represents wakefulness, light sleep, deep sleep and REM sleep. The proprietary algorithm of Fitbit was taken as the baseline model to compare with.

Due to the two-stage nature of our method, there were in total 81 possible combinations because 9 machine learning techniques were available to choose from for each stage. Suppose the combination of machine learning technique i for stage-1 model and technique j for stage-2 model yields an overall performance metric set $\boldsymbol{M}_{ij} = \{Kappa_{ij}, MCC_{ij}\}$, we seek to find the combination of machine learning techniques i and j that achieves the best $\boldsymbol{M}_{ij}$. When we investigated the performance of different combinations, we firstly fixed j (the machine learning technique for stage-2 model) and then identified the best machine learning technique i for stage-1 model that yielded performance $max\ \boldsymbol{M}_j$. We then compared $max\ \boldsymbol{M}_j$ to find the best j.

4 RESULTS

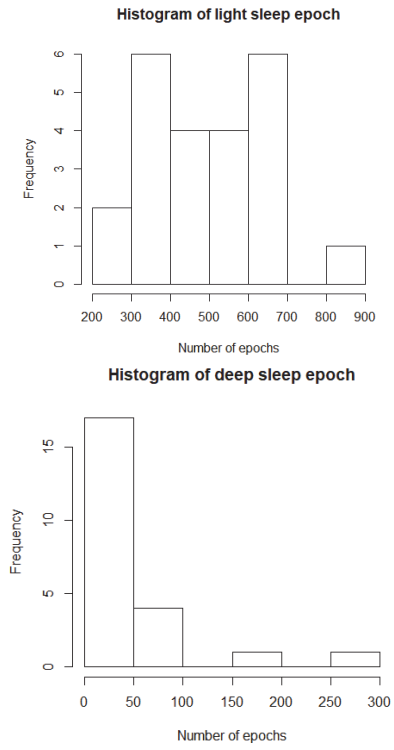

Figure 4: Histogram of light sleep and deep sleep epochs.

Figure 4 and Figure 5 demonstrate the histogram of each sleep stage for the whole cohort, showing that light sleep epochs significantly outnumber deep sleep, REM sleep and wakefulness epochs.

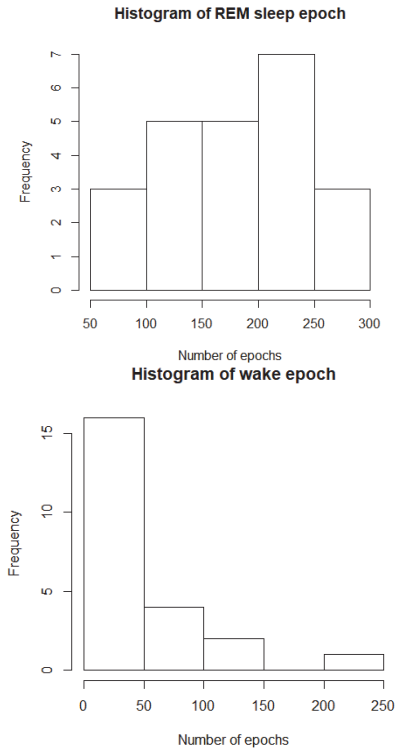

Figure 5: Histogram of REM sleep and wakefulness epochs.

The overall performance of models is summarized in Table 3. The proprietary algorithm of Fitbit was taken as a baseline approach to compare with. The second column in Table 3 shows the best machine learning technique for stage-1 model when the machine learning technique for stage-2 model was fixed. The third and fourth columns are the corresponding *Kappa* and *MCC*. The results show that the overall best combination of machine learning techniques was Naive Bayes for stage-2 model and Naive Bayes with down sampling for stage-1 model. Compared to the baseline method, the best model successfully increased *Kappa* by 27% and *MCC* by 26%. We also found that resampling had no effect on SVM when it was selected as the technique for stage-1 model.

The corresponding accuracy and precision for classifying each sleep stage is demonstrated in Figure 6 and Figure 7 respectively. It is shown that the model that achieved the best overall performance gained advantage mostly by improving the accuracy for light sleep, which was an 29% increase compared to the baseline model. Moreover, the precision for detecting wakefulness and deep sleep was improved by 12% and 57%. Nevertheless, there was no improvement in terms of accuracy and precision for REM sleep, and the accuracy for deep sleep was significantly deteriorated. We also noticed that applying down sampling to Naïve Bayes technique for stage-2 model helped to achieve a more balanced performance for each sleep stage in terms of accuracy and precision.

Table 3: Overall model performance for different combinations of machine learning and resampling techniques compared to baseline model (Fitbit proprietary algorithm).

Stage-2 Model	Best Stage-1 Model	*Kappa*	*MCC*
Fitbit	Fitbit	0.37	0.39
RF	SVM	0.43	0.46
RF-down	SVM	0.43	0.44
RF-up	NB-down	0.43	0.46
NB	NB-down	0.47	0.49
NB-down	SVM	0.40	0.42
NB-up	NB-down	0.42	0.44
SVM	RF	0.44	0.46
SVM-down	SVM	0.40	0.42
SVM-up	SVM	0.40	0.41

5 DISCUSSION

In this study, our model training and testing was conducted in a reliable way through leave-one-out nested cross validation. Resampling was only performed inside cross validation on training set and was not performed on test set to avoid performance bias. Moreover, models were evaluated on a whole night of sleep rather than on a set of random sleep epochs. This ensures that the evaluation results can be generalized to real situations.

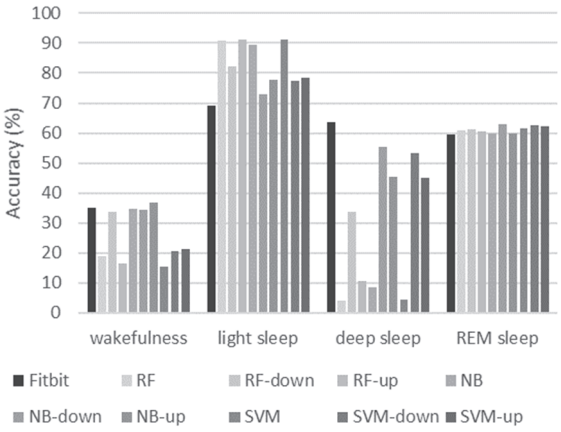

Figure 6: Accuracy for each sleep stage.

Our analysis revealed that it is feasible to improve the accuracy of consumer activity trackers for measuring sleep stages using machine learning techniques. Our proposed two-stage model success-

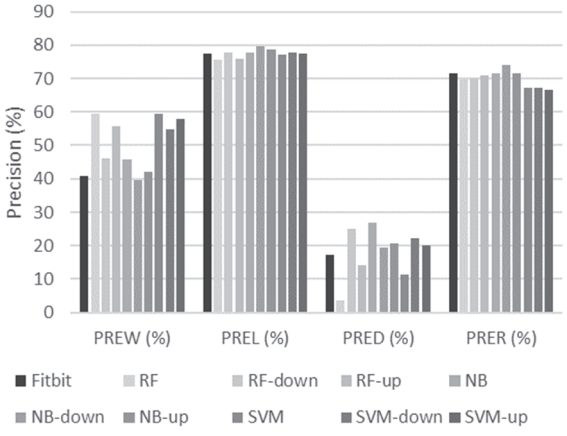

Figure 7: Precision for each sleep stage.

fully improved *Kappa* by up to 27% and *MCC* by up to 26%. As a rule of thumb, a model is generally considered fair if the *Kappa* falls within the range of 0.21-0.40, and moderately good if *Kappa* falls within the range of 0.41-0.60 (J. R. Landis & G. G. Koch, 1977). Based on this scheme, our two-stage model achieved moderately good performance while the baseline model was only fair.

Further examination on the accuracy and precision for each sleep stage revealed that the best performance of the two-stage model was achieved mainly through improving the accuracy for light sleep but at the sacrifice of the accuracy for deep sleep. This may be attributed to the imbalanced nature of the four sleep stages—light sleep tends to be a dominant class. Resampling, especially down sampling, indeed enhanced accuracy on minor classes such as deep sleep and wakefulness. That said, applying resampling had no effect on overall performance when SVM was selected for stage-1 model. Compared to light sleep, other sleep stages were more difficult to classify. The precision for wakefulness and deep sleep were improved at the sacrifice of accuracy. As for REM sleep, the best model could only achieve the same level of accuracy and precision as the baseline model.

The best model was trained using Naive Bayes technique at both stages. Down sampling is required at stage-1 training, but no resampling is needed at stage-2 training. Random forest and support vector machine with linear kernel also achieved good performance but not to the extent of Naive Bayes. These machine learning techniques are computationally effective and have achieved promising results.

6 CONCLUSIONS

Consumer activity trackers have been increasingly used in scientific studies to measure sleep. However, these devices only have limited accuracy for measuring sleep stages. In this study we proposed a two-stage method that post-processed Fitbit data for more accurate sleep stage classification based on machine learning and resampling techniques. The stage-1 model performed binary classification to predict if a Fitbit sleep measurement was correct. The stage-2 model reclassifies incorrect Fitbit sleep measurement into one of the four sleep stages.

Evaluation showed that the best performance was achieved when Naive Bayes was selected as the classification technique at both stages, and random down sampling was applied to stage-1 training. The overall performance metrics *Kappa* and *MCC* were improve by 27% and 26% respectively in the best case. Further examination on the accuracy and precision for each sleep stage revealed that the best performance of the two-stage model was achieved mainly through improving the accuracy for light sleep but at the sacrifice of the accuracy for deep sleep. Resampling techniques, especially down sampling, helped to enhance accuracy for minor classes such as deep sleep and wakefulness.

The proposed method can be used to post-process Fitbit data for more accurate measurement of sleep stages. It can also be easily adapted to other off-the-shelf wearable activity trackers that are based on similar mechanism. The outcome of this study can benefit researchers who intend to use these devices in scientific studies and individual users who rely to Fitbit data to make self-care decisions. In the next step, we will investigate the performance of more advanced machine learning techniques such as XGBoost. We will also use more advanced techniques such as segmented modelling to address the class imbalance issue.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number 16H07469 and 19K20141. The authors would like to thank the participants of the data collection experiment.

REFERENCES

Ancoli-Israel, I., Chesson, A., & Quan, S. (2017). The AASM manual for the scoring of sleep and associated events rules, terminology and technical specifications. *Darien, IL: American Academy of Sleep Medicine, Version 2.4*.

Bian, J., Guo, Y., Xie, M., Parish, A., at al. (2017). Exploring the association between self-reported asthma impact and Fitbit-derived sleep quality and physical activity measures in adolescents. *JMIR mHealth and uHealth*, 5(7), e105.

Buysse, D., Reynolds, C., Monk, T., at al. (1989). The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry Res*, 28(2), 193-213.

Colten, H. R., & Altevogt, B. M. (2006). *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*. Washington, DC: National Academies Press.

De Zambotti, M., Cellini, N., Goldstone, A., Colrain, I., & Baker, F. (2019). Wearable sleep technology in clinical and research settings. *Med Sci Sports Exerc*, 51(7), 1538-1557.

De Zambotti, M., Goldstone, A., Claudatos, S., at al. (2017). A validation study of Fitbit Charge 2 compared with polysomnography in adults. *Chronobiology International*, 35(4), 465-476.

Goldstone, A., Baker, F. C., & De Zambotti, M. (2018). Actigraphy in the digital health revolution: still asleep? *Sleep*, 41(9), zsy120.

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, 28(5), 367-374.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Liang, Z., & Chapa-Martell, M. A. (2018a). Not all errors are created equal: influence of user characteristics on measuring errors of consumer wearable devices for sleep tracking. *EAI Endorsed Transactions on Pervasive Health and Technology*, 18(15), e4.

Liang, Z., & Chapa-Martell, M. A. (2018b). Validity of consumer activity wristbands and wearable EEG for measuring overall sleep parameters and sleep structure in free-living conditions. *Journal of Healthcare Informatics Research,* 1-27.

Liang, Z., & Chapa-Martell, M. A. (2019a). Accuracy of Fitbit wristbands in measuring sleep stage transitions and the effect of user-specific factors. *JMIR Mhealth Uhealth*, 7(6), e13384.

Liang, Z., & Chapa-Martell, M. A. (2019b). *Achieving accurate ubiquitous sleep sensing with consumer wearable activity wristbands using multi-class imbalanced classification*. Paper presented at the PICOM 2019.

Liang, Z., & Chapa-Martell, M. A. (2019c). *Combining numerical and visual approaches in validating sleep data quality of consumer wearable wristbands*. Paper presented at the PerCom Workshops.

Liang, Z., & Chapa-Martell, M. A. (2019d). *Combining resampling and machine learning to improve sleep-wake detection of Fitbit wristbands*. Paper presented at ICHI 2019.

Liang, Z., Chapa-Martell, M. A., & Nishimura, T. (2016a). *A personalized approach for detecting unusual sleep from time series sleep-tracking data*. Paper presented at ICHI 2016.

Liang, Z., Chapa-Martell, M. A., & Nishimura, T. (2016b). *Mining hidden correlations between sleep and lifestyle factors from quantified-self data*. Paper presented at the UbiComp 2016.

Liang, Z., Ploderer, B., Liu, W., et al. (2016). SleepExplorer: a visualization tool to make sense of correlations between personal sleep data and contextual factors. *Personal Ubiquitous Comput*.

Liang, Z., & Ploderer, B. C.-M., Mario Alberto. (2017). *Is fitbit fit for sleep-tracking?: sources of measurement errors and proposed countermeasures*. Paper presented at PervasiveHealth 2017.

Liu, W., Ploderer, B., & Hoang, T. (2015). *In Bed with Technology: Challenges and Opportunities for Sleep Tracking*. Paper presented at OzCHI 2015.

Meltzer, L., Hiruma, L., Avis, K., at al. (2015). Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. *Sleep*, 38(8), 1323-1330.

Nam, Y., Kim, Y., & Lee, J. (2016). Sleep monitoring based on a tri-axial accelerometer and a pressure sensor. *Sensors*, 16(5), 750.

Rahman, T., Adams, A., Ravichandran, R. V., et al. (2015). *DoppleSleep: a contactless unobtrusive sleep sensing system using short-range Doppler radar.* Paper presented at UbiComp 2015.

Shelgikar, A., Anderson, P., & Stephens, M. (2016). Sleep tracking, wearable technology, and opportunities for research and clinical care. *CHEST*, 150(3), 732-743.

Weatherall, J., Paprocki, Y., Meyer, T. M., et al. (2018). Sleep tracking and exercise in patients with type 2 diabetes mellitus (step-D): pilot study to determine correlations between Fitbit data and patient-reported outcomes. *JMIR Mhealth Uhealth*, 6(6), e131.

Weaver, G., Beets, M., Perry, M., et al. (2018). Changes in children's sleep and physical activity during a one-week versus a three-week break from school: a natural experiment. *Sleep*, zsy205.

Xiao, M., Yan, H., Song, J., et al. (2013). Sleep stage classification based on heart rate variability and random forest. *Biomedical Signal Processing and Control*, 8, 624-633.

Zhu, G., Li, Y., & Wen, P. (2014). Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal. *IEEE Journal of Biomedical and Health Informatics,* 18(6), 1813 - 1821.