Predicting Medical-Grade Sleep-Wake Classification from Fitbit Data Using Tree-Based Machine Learning

ZILU LIANG[†1,†2] MARIO ALBERTO CHAPA-MARTELL[†3]

Abstract: It has become increasingly popular among individuals and researchers to monitor sleep using consumer activity wristbands. Nevertheless, many validation studies have identified a significant gap between consumer wristbands and medical sleep monitors. This study aims to bridge this gap through developing predictive models that leverage Fitbit data to generate medical-grade sleep/wake classification. Considering that the "sleep" class significantly outnumbers the "wake" class, we formulated the problem of interest into an imbalanced classification problem. We applied two tree-based machine learning techniques, i.e. decision tree and random forest, in combination with four re-sampling methods, which yields in total 10 classifiers. The performance of the classifiers was compared to the original Fitbit algorithm based on sensitivity, specificity and area under the ROC curve (AUC). Our results showed that in the best case, specificity was improved by 75% while sensitivity was reduced by 12%, which yielded a statistically significant increase of 11% in AUC. The decision tree technique was more robust and less affected by re-sampling method compared to the random forest technique, and random up sampling may be a most effective re-sampling strategy to balance the training sets. These findings demonstrate the feasibility of achieving medical-grade sleep/wake classification from consumer wristbands by applying proper combination of reesampling and machine techniques.

Keywords: Consumer wearable; Fitbit; Sleep; Machine learning; Decision tree; Random forest; Imbalanced classification

1. Introduction

Consumer sleep tracking devices are becoming more and more popular not only among individual users who are curious about their sleep patterns but also among researchers who intend to collect longitudinal data ecologically [1-5]. Nevertheless, these devices are not able to achieve medical-grade accuracy [1, 2, 6-9]. Many validation studies have endeavored to establish the discrepancy between consumer wristbands and medical sleep monitors. There is strong evidence that previous models of consumer wristbands such as Fitbit Flex and Fitbit Charge overestimated sleep time while underestimated wake time [5, 10-13]. This problem can be traced back to the intrinsic limitation of medical actigraphy which consumer wristbands share the same mechanism with [13-17]. Recent models such as Fitbit Charge 2 use multiple streams of bio-signals including locomotion and heart rate to infer sleep stages, but their accuracy is still not satisfactory [18, 19]. A large body of research has been devoted to developing brand new sleep trackers for better accuracy. Nevertheless, it is not likely that these new devices will gain a larger user base worldwide than well-established manufacturers such as Fitbit in the near future. Therefore, re-engineering data from popular consumer wristbands for better accuracy could potentially make a stronger impact than making new devices.

The basic idea of our proposal is to train a classification model that takes Fitbit data and demographic information as input to predict medical-grade output. Since the "sleep" class outnumbers the "wake" class by a large proportion, the problem of interest was formulated into an imbalanced binomial classification problem. Four re-sampling methods are applied to the dataset to generate balanced training sets: random up sampling, random down sampling, random over-sampling examples (ROSE), and synthetic minority oversampling technique (SMOTE). We applied tree-based machine learning techniques including decision tree and random forest because they can handle both continuous and categorical features. Previous studies have demonstrated their merit in solving classification problems in sleep research [20, 21]. The performance of the classification models is evaluated based on sensitivity (indicating the ability of a classifier in detecting sleep epochs), specificity (indicating the ability of a classifier in detecting wake epochs), and AUC (indicating the overall classification performance).

The contribution of this study is two-fold. First, we proposed a promising solution to improve the overall performance of Fitbit (and wristband-type sleep trackers in general) in sleep/wake classification, and especially in terms of specificity. Second, we examined the performance of different combinations of classification technique and re-sampling methods. The results produce rich implications to future research along the same line. The rest of the papers is organized as follows. Section 2 summarizes related work on sleep scoring in clinical settings and the validity of consumer sleep tracking technologies. Section 3 and 4 present the proposed methodology and the performance evaluation. We discuss the strength and weakness of different classifiers in Section 5 and close the whole paper in the conclusions.

2. Related Work

In recent years many sleep tracking wristbands appeared in the consumer market [1, 2], featuring popular brands such as Fitbit, Apple, Garmin, etc. These consumer sleep tracking technologies resemble clinical actigraphy that infer sleep/wake based on a person's movement [17] and they offer an ecological method for individuals to monitor their sleep at home. Fitbit devices offer two working modes: the "normal" mode for

†1 Kyoto University of Advanced Science
†2 The University of Tokyo
†3 CAC Corporation

healthy users and the “sensitive” mode for people with sleep disorders. Many studies show that these devices have positive impact on people’s sleep hygiene and raise people’s awareness of sleep health [3, 22]. In the meanwhile, there are increased number of research studies using consumer sleep tracking devices to measure sleep outcomes [3, 23].

Nevertheless, the measurement accuracy of consumer sleep trackers has raised wide concern [2, 9]. Validation studies on older models of consumer sleep-tracking wristbands found that these devices overestimated sleep while underestimated wake [5, 10-12]. Epoch-wise comparison between Fitbit and PSG demonstrated high sensitivity (i.e. the accuracy in classifying sleep, range: 80% ~ 96%) but low specificity (i.e. the accuracy in classifying wake, range: 40% ~ 61%) when used in normal mode [5, 18, 19, 24-26]. When used in sensitive mode, the specificity may be improved at the sacrifice of reducing sensitivity [10, 12, 27]. To this end, relying solely on locomotion and heart rate has been considered insufficient for accurate classification of sleep and wake. In this study, we adopt a novel approach that leverages Fitbit data to predict medical-grade sleep-wake classification.

3. Methodology

3.1 Data collection

Sleep data is simultaneously collected using Fitbit Charge 2 device and a medical device. The PSQI (Pittsburgh Sleep Quality Index) [28] is used to measure subjective sleep quality as well as recording basic demographic information such as sex and age.

The Fitbit Charge 2 wristbands use imbedded optical sensors and accelerometers to measure heart rate and locomotion respectively. These data are then used to infer other physiological and behavioral data such as steps, exercise, heart rate, and sleep. The processed data are presented to users on a dashboard on the Fitbit smartphone application. In this study, we take advantage of the sleep data (both aggregate sleep data and intra-day sleep data) and heart rate data (intra-day heart rate data during sleep) from the Fitbit devices. We extracted aggregate sleep data through the Fitbit public API using a web application that we developed in our previous study [3]. As for the intra-day sleep and heart rate data, we extracted the data at 1s resolution through the Fitbit partner API upon getting permissions from the Fitbit Company.

Different from Fitbit devices that rely on accelerometer and optical sensors, the medical device, i.e. Sleep Scope, is a single channel EEG that measures brainwaves. The Sleep Scope device has been validated against the golden standard PSG [29][30]. The main body of the device is connected to two gel-type electrodes by cables. Users need to stick one of the electrodes on their forehead and the other behind an ear. The raw EEG data was analyzed by the company, firstly using proprietary auto-scoring software and then inspected and revised by sleep experts based on sleep scoring standards [31].

3.2 Feature Construction

TABLE I. FULL LIST OF FEATURES

Level	Feature	Type (Unit)	Measuring Method
Macro-level	Sex	Nominal (0=female/1=male)	Self-report
	Age	Ordinal	Self-report
	PSQI	Ordinal	PSQI [28]
	Total sleep time (TST)	Continuous (s)	Fitbit Charge 2
	Wake after sleep onset (WASO)	Continuous (s)	Fitbit Charge 2
	Sleep efficiency (SE)	Continuous	Fitbit Charge 2
	Wake ratio	Continuous	Fitbit Charge 2
	Light sleep ratio	Continuous	Fitbit Charge 2
	Deep sleep ratio	Continuous	Fitbit Charge 2
	REM sleep ratio	Continuous	Fitbit Charge 2
Micro-level	k	Ordinal	Fitbit Charge 2
	Sleep(k)	Nominal (0=wake/1=sleep)	Fitbit Charge 2
	HR (k)	Ordinal	Fitbit Charge 2
	$\frac{\mathrm{HR}(k)-\mathrm{HR}(k-1)}{\mathrm{HR}(k-1)}$	Continuous	Fitbit Charge 2

The input data that we use to construct features include Fitbit data (both daily aggregate data and intra-day data), demographic information (i.e. sex and age), and subjective sleep quality measured by the PSQI [28]. We extract two types of features: macro-level features and micro-level features. The macro-level features include sex, age, PSQI, and daily aggregate sleep data (i.e. total sleep time, wake after sleep onset, sleep efficiency, wake ratio, light sleep ratio, deep sleep ratio, and REM sleep ratio). These macro-level features do not vary across the samples from a certain participant. The micro-level features include intra-day sleep and heart rate averaged every 30s epoch, the change in heart rate compared to previous epoch, and the epoch ID. Each epoch corresponds to an instance in the dataset. The epoch ID captures the temporal information of an instance. This is important as human sleep demonstrates temporal patterns [32]. The medical data are synchronized with the Fitbit data, aggregated into 30s epoch, and are used as the labels for each corresponding instance. All the features are listed in Table I, where k, Sleep(k) and HR(k) denote the k-th epoch, the average sleep status in epoch k, and the average heart rate in epoch k. In this study, k is within the range of 1~1208, Sleep(k) is either 0 (indicating wake) or 1 (indicating sleep).

3.3 Training set preprocessing

One characteristic of human sleep is that the "sleep" epochs significantly outnumbers "wake" epochs. Applying standard machine learning techniques directly to the imbalanced training sets will likely lead to classification models that are biased towards the "sleep" class. To mitigate this problem, we applied re-sampling strategies list below [33].

- *Random up sampling* [34] randomly generates artificial instances to the "wake" class so that the frequency of the "wake" class is close to that of the "sleep" class.
- *Random down sampling* [34] randomly subsets the "sleep" class to match the frequency of the "wake" class.
- *Random over-sampling examples (ROSE)* [35] generates new artificial samples to the "wake" class according to a smoothed bootstrap approach that combines techniques of up sampling and down sampling.
- *Synthetic minority oversampling technique (SMOTE)* [36] synthesizes artificial data in the "wake" class based on the feature space similarities between existing "wake" samples.

3.4 Model Training, Tuning and Testing

Two tree-based machine learning techniques are applied to achieve the classification purpose. The decision tree technique relies on a recursive partitioning strategy that repeatedly partitions the predictor space into multiple simple regions, so that the outcomes in each final subset is as homogeneous as possible [37]. This process generates a set of splitting rules that can be used to classify new data. In this study, information gain (i.e. entropy) was used as the split measure.

TABLE II. DENOTATION OF ALL CLASSIFIERS

	Decision Tree Classifiers	Random Forest Classifiers
Original imbalanced training sets	DT	RF
Randomly up sampled training sets	DT-U	RF-U
Randomly down sampled training sets	DT-D	RF-D
ROSE sampled training sets	DT-R	RF-R
SMOTE sampled training sets	DT-S	RF-S

The random forest (RF) technique is an ensemble learning method that combines many decision trees to yield a single consensus prediction [38]. In this study, each tree of the random forest is built using the CART method without pruning, and 500 trees were used. Random forest is a more robust technique compared to single decision trees and it has the advantages of fast computation, high accuracy, and robust to noise compared to other machine learning techniques. Several studies have proposed to use random forest for automatic sleep scoring [39, 40]. Combining the two techniques with different re-sampling strategies listed up in the previous subsection, we obtain in total 10 classifiers summarized in Table II.

We use a leave-one-out strategy for validating the performance of the classification models. In each iteration, the data of participant *n*, i.e. Dataset *n*, is used as the test set, while the data of all other *N*-1 participants (*N* is the total number of participants; in this study, *N* = 23) are merged into one large dataset as the training set. As is illustrated in Fig. 1, this process is iterated *N* times and the average values across the *N* iterations are used as the final results. In each iteration, the classification model is trained using 10-fold cross validation with 3 repeated times to avoid overfitting [41, 42]. The parameters in the classification models are tuned using random search with tune length = 8 to overcome any biases of manual tuning. Paired *t*-test is conducted to examine if there is any statistically significant differences between the predictions made by the classification models and those made by the proprietary Fitbit algorithm.

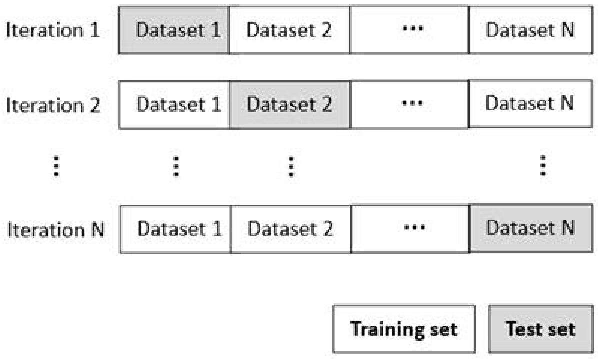

Fig. 1. Leave-one-out cross validation.

4. Evaluation

4.1 Performance measures

TABLE I. PERFORMANCE MEASURES

Measure	Definition	Interpretation
Sensitivity	$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$	The proportion of true positive that are predicted as positive, also called recall.
Specificity	$\frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$	The proportion of true negative that are predicted as negative.
AUC	The area under the ROC curve	A general measure of predictiveness.

The performance of the classification models is evaluated using the measures summarized in Table III, where TP, TN, FP, FN denote true positive (i.e. the number of sleep epochs that are correctly classified as sleep), true negative (i.e. the number of wake epochs that are correctly classified as wake), false positive (i.e. the number of wake epochs that are incorrectly classified as sleep), and false negative (i.e. the number of sleep epochs that are incorrectly classified as wake). Sensitivity indicates the accuracy of a classification model in detecting sleep epochs, and specificity indicates the accuracy in detecting wake epochs [43]. The AUC (Area Under the ROC Curve) provides an aggregate measure of a classifier's performance on average [44, 45]. Other metrics such as accuracy and F1 score are not used as they tend to be deceiving in the case of extremely imbalanced datasets [46]. The machine learning process and statistical analysis were conducted in open source software R [47].

4.2 Descriptive statistics of datasets

We collected data from 23 healthy adults (9 female, age range: 21 ~ 30 years). The total number of epochs from the sleep hypnogram of each participant ranges between 418 ~ 1208 (on average 777). The distribution of the number of sleep epochs and that of the number of wake epochs are shown in Fig. 2 and Fig. 3. The number of sleep epochs demonstrates a binomial distribution with a peak between 600 ~ 1000 epochs, while the number of wake epochs demonstrates a Poisson distribution with a peak between 0 ~ 50 epochs. In each iteration of the leave-one-out test, the training set consists of 16671 ~ 17461 samples and the test set consists of 418 ~ 1208 samples. The sample size is sufficient as suggested by previous studies [48].

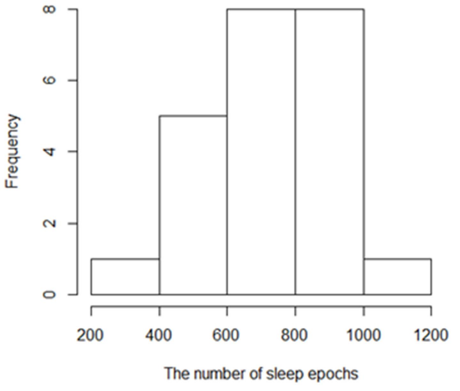

Fig. 2. Distribution of the positive class (i.e. the sleep class).

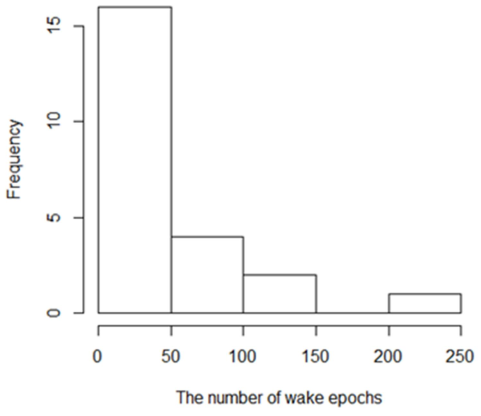

Fig. 3. Distribution of the negative class wake (i.e. the wake class).

A. Classification Performance

We conducted epoch-wise comparison between the predicted values and the true values to calculate the performance evaluation measures. The average performance of all models is summarized in Table IV. Asterisks indicate statistically significant differences to the proprietary Fitbit algorithm. The results show that in the best case, specificity was improved by 75% while sensitivity was reduced by 12%, which yielded a statistically significant increase of 11% in AUC.

We also used box-and-whisker plots to demonstrate the distribution (i.e. minimum, first quartile, median, third quartile, and maximum) of the performance measures for all classification models [49]. As shown in Fig. 4 ~ Fig. 6, the white, light grey and dark grey boxes indicate the performance of the proprietary Fitbit algorithm, the decision tree classifiers, and the random forest classifiers respectively. The data points that fall outside the maximum and minimum range are outliers.

TABLE II. AVERAGE PERFORMANCE OF ALL MODELS

	Sensitivity (%)	Specificity (%)	AUC
Fitbit	96.4±2.4[a]	35.0±19.5	65.7±9.7
DT	97.1±4.0	17.1±11.6	57.5±5.0**[b,c]
DT-U	84.8±13.8***	61.4±16.7***	73.1±7.2**
DT-D	85.3±14.8**	58.8±16.8***	72.0±7.9*
DT-R	83.6±11.9***	59.9±19.5***	71.8±8.9*
DT-S	88.4±9.7***	55.6±20.6**	72.0±8.9*
RF	97.4±3.6	16.4±12.3***	57.1±5.2***
RF-U	83.0±16.7**	58.9±20.7***	71.0±7.3*
RF-D	96.0±4.2	16.4±14.5***	56.8±5.6***
RF-R	78.5±21.6***	55.8±23.8**	67.2±8.6
RF-S	86.4±14.5**	36.7±23.6	63.4±8.5

a. The results are presented in the form of "average ± standard deviation".

b. Statistical significance is based on comparison to the proprietary Fitbit algorithm.

c. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

Fig. 4 shows that the sensitivity of the classification models was scattered with many outliers. Paired *t*-test shown in Table IV demonstrated that classifiers DT, RF, RF-D were not significantly different from the proprietary Fitbit algorithm. The average sensitivity of all other classifiers were lower than that of the Fitbit. Among the decision tree classifiers, the ones that were trained using re-sampled data yielded significantly lower sensitivity compared to DT. In the meanwhile, different re-sampling method did not affect the sensitivity of decision tree classifiers, as no statistically significant difference was found among the average sensitivity of DT-U, DT-D, DT-R, and DT-S. Among the random forest classifiers, the ones that were trained using re-sampled data also yielded significantly lower sensitivity compared to RF, except RF-D. No statistically significant difference was found in terms of average sensitivity among RF-U, RF-R, and RF-S.

With respect to specificity, Fig. 5 demonstrates significant improvement of decision tree classifiers with all re-sampling methods and random forest classifiers with random up sampling and ROSE. The average specificity of classifier RF-S was statistically the same as that of the proprietary Fitbit algorithm, whereas classifiers DT, RF and RF-D had worse performance compared to Fitbit, all with strong statistical significance ($p < 0.001$). All decision tree classifiers with re-sampled datasets had significantly better specificity compared to Fitbit and DT. Moreover, no statistically significant difference was found among these classifiers. Among the random forest classifiers, RF-U, RFR and RF-S yielded higher average specificity than Fitbit and RF. In addition, the average specificity of RF-S was significantly lower than that of RF-U ($p = 0.002$) and RF-R ($p = 0.010$).

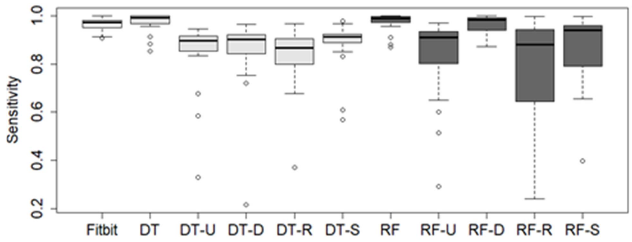

Fig. 4. Box-and-whisker plots of sensitivity.

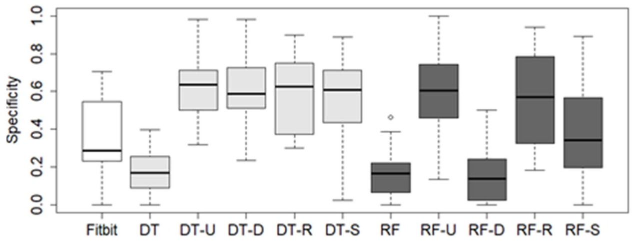

Fig. 5. Box-and-whisker plots of specificity.

The overall performance of the classifiers indicated by the AUC is shown in Fig. 6. Classifiers DT-U ($p = 0.007$), DT-D ($p = 0.023$), DT-R ($p = 0.039$), DT-S ($p = 0.031$), RFU ($p = 0.05$) all had statistically significant improvement compared to the baseline Fitbit algorithm. Classifiers RF-R and RF-S had equivalent performance as Fitbit, whereas classifiers DT ($p = 0.001$), RF ($p < 0.001$) and RF-D ($p < 0.001$) had worse performance compared to the baseline Fitbit algorithm. All decision tree classifiers with re-sampled training data demonstrated statistically equivalent performance with each other, which was better than the original decision tree classifier DT. The AUC of RF-S was significantly lower than that of RF-U ($p = 0.003$), but was statistically equivalent with that of RF-R.

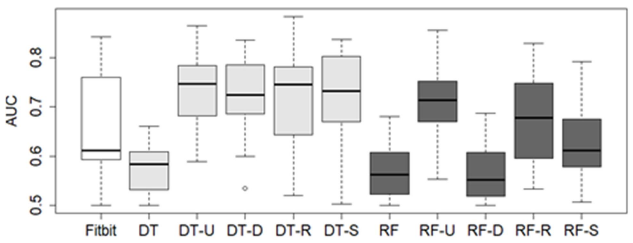

Fig. 6. Box-and-whisker plots of AUC.

5. Discussion

The purpose of this study was to develop predictive models that leverage Fitbit data to generate medical-grade sleep/wake classification. The problem of interest differs from standard binomial classification problems in that the "sleep" class outnumbers the "wake" class by a magnitude of 10 ~ 100. Therefore, applying machine learning techniques without addressing the imbalanced datasets was inadequate for building good classification models. As demonstrated by the evaluation results, the original decision tree classifier (DT) and the original random forest classifier (RF) were indeed biased towards the majority class (i.e. high sensitivity and low specificity).

Evaluation also showed that proper re-sampling method could significantly improve a classifier's overall performance as indicated by the AUC. Nevertheless, different combinations of machine learning technique and re-sampling strategy may yield distinct performance. Our analysis showed that there was always trade-off between sensitivity and specificity. Better specificity was always achieved at the sacrifice of sensitivity, though reduced sensitivity not necessarily corresponds to enhanced specificity. In what follows, we discuss the strengths and weaknesses of different re-sampling method and machine learning technique as well as the limitations of this study.

5.1 Comparison of Re-sampling Methods

Re-sampling has been a popular strategy for addressing imbalance data [46]. We have examined the effect of four different re-sampling methods on the overall performance of the classifiers. Our results showed that in general up sampling yielded better performance than down sampling. Random up sampling and ROSE produced consistently good performance regardless of the machine learning technique applied, and the former demonstrated less variance compared to the latter. To this end, the effectiveness of up sampling methods also echoes findings in previous studies in neurocomputing and finance that up sampling methods were superior to down sampling methods when there were only a few dozen minority instance, and vice versa when there were hundreds of minority samples [50, 51].

Although it has been widely recognized that multiplicities in random up sampling may become "tied" and thus leading to overfitting, this did not occur in this study probably due to the relatively small sample size corresponding to each epoch ID. The synthesized up sampling method SMOTE enhanced the sensitivity of the random forest classifier towards the minority class (i.e. wake) compared to its counterpart trained using imbalanced dataset, but not to the extent of outperforming the original Fitbit algorithm with statistical significance. One possibility is that the synthesize process modified the distribution in terms of the feature "epoch ID", whereas the other synthesized method ROSE drew new examples from an estimate of the conditional density underlying the data and thus ensured the distribution of the data into the class was not changed.

As for the down sampling method, previous studies found that it may lead to information loss and thus worsen the sensitivity to the majority class [34]. Nevertheless, this is not the case in this study as the sensitivity of the classifier was not significantly reduced after re-sampling. On the contrary, we hypothesize that the deteriorated performance of the random down sampling method on specificity may be caused by the small sample size after re-sampling.

5.2 Comparison of Classification Techniques

In machine learning, tree-based techniques have several advantages in terms of scalability and robustness to outliers. Random forest has been considered more robust than simple decision tree as it is an ensemble method that combines predictions from many individual trees. Nevertheless, our analysis showed that simple decision tree classifiers consistently outperformed random forest classifiers regardless of the data re-sampling method applied.

The performance of random forest classifiers was overall mediocre. On one hand, up sampling did not improve the overall performance of random forest classifiers. Despite of the enhanced performance in detecting wake (i.e. better specificity), the performance in detecting sleep (i.e. sensitivity) was significantly reduced, thus worsening the overall predictiveness of the model as indicated by AUC. On the other hand, random forest with down sampled training data has deteriorated performance in terms of specificity and AUC compared to the proprietary Fitbit algorithm. This indicates that the random forest technique may be more sensitive to sparse data due to down sampling.

This result may contradict the impression that random forest achieves better performance than simple decision tree due to the ensemble process. However, several studies have found that random forest may not be suited for time series classification as it is not able to capture the temporal information of the dataset. In this study, the temporal characteristics of the sleep hypnogram was to a large extent erased as we mapped 30-s epochs to individual samples. Nevertheless, the "epoch ID" feature still incorporated sequential information, which may be the reason for the unsatisfactory performance of the random forest classifiers.

5.3 Limitations

This study has the following limitations. First, we did not examine the performance of the classification models on the aggregate level. Given the small portion of wake epochs, it remains unclear to what extent the benefit of enhanced specificity can be translate into enhanced measurement accuracy of total sleep time and total wake time. Second, we did not investigate the correction power of the classification models, i.e. to what extent the classification model corrected the classification errors of the proprietary Fitbit algorithm. Third, the methods that we applied to mitigate imbalance data were not exhaustive. We only investigated the performance of a few data-level methods. Other re-sampling strategies such as adaptive multiple re-sampling [52], boosting based synthetic over-sampling [53] and other ensemble methods [54] should be examined in future studies. In addition, future research may also apply algorithm-level approaches [55] and cost-sensitive learning [56].

6. Conclusion

We have proposed and evaluated a machine learning based method for predicting medical-grade sleep/wake classification from Fitbit data. We investigated the performance of the classification models combining different machine learning techniques (i.e. decision tree and random forest) and re-sampling methods (i.e. random up sampling, random down sampling, ROSE, and SMOTE). Our results showed that in the best case, specificity was improved by 75% while sensitivity was reduced by 12%, which yielded a statistically significant increase of 11% in AUC. Evaluation also showed that up sampling methods yielded better performance than down sampling method, and decision tree consistently outperformed random forest regardless of the re-sampling method applied. We conclude that up sampling combined with decision tree may be most suited for the problem of interest.

Acknowledgments This work was supported by JSPS KAKENHI Grant Number 16H07469 and 19K20141.

Reference

[1] M. Bianchi, "Sleep devices: wearables and nearables, informational and interventional, consumer and clinical," Metabolism, vol. 84, pp. 99-108, 2018.

[2] K. G. Baron, J. Duffecy, M. A. Berendsen, I. C. Mason, E. G. Lattie, and N. C. Manalo, "Feeling validated yet? A scoping review of the use of consumer-targeted wearable and mobile technology to measure and improve sleep," Sleep Medicine Reviews, vol. 40, pp. 151-159, 2018.

[3] Z. Liang, B. Ploderer, W. Liu, Y. Nagata, J. Bailey, L. Kulik, and Y. Li, "SleepExplorer: a visualization tool to make sense of correlations between personal sleep data and contextual factors," in Personal Ubiquitous Comput., 2016, pp. 985-1000.

[4] Z. Liang, B. Ploderer, M. A. Chapa-Martell, and T. Nishimura, "A cloud-based intelligent computing system for contextual exploration on personal sleep-tracking data using association rule mining," Intelligent Computing Systems. Communications in Computer and Information Science, A. Martin-Gonzalez and V. Uc-Cetina, eds.: Springer, Cham, 2016.

[5] M. De Zambotti, F. Baker, C, and A. R. Willoughby, "Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents," Physiological & Behavior, vol. 158, pp. 143-149, 2016.

[6] L. Piwek, D. A. Ellis, S. Andrews, and A. Joinson, "The rise of consumer health wearables: promises and barriers," PLoS Med, vol.

13, no. 2, pp. e1001953, 2016.

[7] B. Kolla, S. Mansukhani, and M. Mansukhani, “Consumer sleep tracking devices: a review of mechanisms, validity and utility,” Expert Review of Medical Devices, vol. 13, no. 5, pp. 497-506, 2016.

[8] J. M. Peake, G. Kerr, and J. P. Sullivan, “A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations,” Frontiers in Physiology, vol. 9, pp. 743, 2018.

[9] Z. Liang, and B. C.-M. Ploderer, Mario Alberto, “Is fitbit fit for sleep-tracking?: sources of measurement errors and proposed countermeasures,” in Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, Barcelona, Spain, 2017, pp. 476-479.

[10] J. Cook, M. Prairie, and D. Plante, “Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: a comparison against polysomnography and wrist-worn actigraphy,” J Affect Disorder, vol. 217, pp. 299-305, 2017.

[11] S.-G. Kang, J. M. Kang, K.-P. Ko, P. Seon-Cheol, S. Mariani, and J. Weng, “Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers,” Journal of Psychosomatic Research, vol. 97, pp. 38-44, 2017.

[12] L. Meltzer, L. Hiruma, K. Avis, and e. al., “Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents,” Sleep, vol. 38, no. 8, pp. 1323-1330, 2015.

[13] V. Natale, G. Plazzi, and M. Martoni, “Actigraphy in the assessment of insomnia: a quantitative approach,” Sleep, vol. 32, no. 6, pp. 767-771, 2009.

[14] T. Blackwell, S. Ancoli-Israel, S. Redline, and K. Stone, “Factors that may influence the classification of sleep-wake by wrist actigraphy: the MrOS sleep study,” J Clin Sleep Med, vol. 7, no. 4, pp. 357-367, 2011.

[15] J. Martin, and A. Hakim, “wrist actigraphy,” CHEST, vol. 139, no. 6, pp. 1514-1527, 2011.

[16] M. Marino, Y. Li, M. Rueschman, and e. al., “Measuring sleep accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography,” Sleep, vol. 36, no. 11, pp. 1747-1755, 2013.

[17] V. Natale, D. Leger, M. Martoni, and e. al., “The role of actigraphy in the assessment of primary insomnia: a retrospective study,” Sleep Medicine, vol. 15, no. 1, pp. 111-115, 2014.

[18] M. De Zambotti, A. Goldstone, S. Claudatos, and e. al., “A validation study of Fitbit Charge 2 compared with polysomnography in adults,” Chronobiology International, vol. 35, no. 4, pp. 465-476, 2017.

[19] Z. Liang, and M. A. Chapa-Martell, “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, pp. 1-27, 2018.

[20] S.-F. Liang, C.-E. Kuo, Y.-H. Hu, and Y.-S. Cheng, “A rule-based automatic sleep staging method,” Journal of Neuroscience Methods, vol. 205, pp. 169-176, 2012.

[21] T. Lajnef, S. Chaibi, P. Ruby, A. Pierre-Emmanuel, E. Jean-Baptiste, M. Samet, A. Kachouri, and K. Jerbi, “Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines,” Journal of Neuroscience Methods, vol. 250, pp. 94-105, 2015.

[22] W. Liu, B. Ploderer, and T. Hoang, “In Bed with Technology: Challenges and Opportunities for Sleep Tracking,” in Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction, Parkville, VIC, Australia, 2015, pp. 142-151.

[23] G. Bellone, S. Plano, D. Cardinali, D. Perez Chada, D. Vigo, and D. Golombek, “Comparative analysis of actigraphy performance in healthy young subjects,” Sleep Science, vol. 9, pp. 272-279, 2016.

[24] M. De Zambotti, F. C. Baker, and I. M. Colrain, “Validation of sleep-tracking technology compared with polysomnography in adolescents,” Sleep, vol. 38, no. 9, pp. 1461-1468, 2015.

[25] M. De Zambotti, S. Claudatos, S. Inkelis, I. Colrain, and F. Baker, “Evaluation of a consumer fitness-tracking device to assess sleep in adults,” Chronobiology International, vol. 32, no. 7, pp. 1024-1028, 2015.

[26] A. Goldstone, F. C. Baker, and M. De Zambotti, “Actigraphy in the digital health revolution: still asleep?,” Sleep, vol. 41, no. 9, pp. zsy120, 2018.

[27] E. Toon, M. Davey, S. L. Hollis, G. M. Nixon, R. S. Horne, and S. N. Biggs, “Comparison of commercial wrist-based and smartphone accelerometers, a ctigraphy, and PSG in a clinical cohort of children and adolescents,” Journal of Clinical Sleep Medicine, vol. 12, no. 3, 2016.

[28] D. Buysse, C. Reynolds, T. Monk, and e. al., “The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research,” Psychiatry Res, vol. 28, no. 2, pp. 193-213, 1989.

[29] M. Yoshida, H. Shinohara, and H. Kodama, “Assessment of nocturnal sleep architecture by actigraphy and one-channel electroencephalography in early infancy,” Early Human Development, vol. 91, no. 9, pp. 519-526, 2015.

[30] R. Rosenberg, and S. Van Hout, “The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring,” J Clin Sleep Med, vol. 9, no. 1, pp. 81-87, 2013.

[31] I. Ancoli-Israel, A. Chesson, and S. Quan, “for the American Academy of Sleep Medicine. The AASM manual for the scoring of sleep and associated events rules, terminology and technical specifications,” Darien, IL: American Academy of Sleep Medicine, pp. Version 2.4, 2017.

[32] M. Ohayon, E. M. Wickwire, M. Hirshkowitz, and e. al., “National Sleep Foundation's sleep quality recommendations: first report,” Sleep Health, vol. 3, no. 1, pp. 6-19, 2017.

[33] B. Krawczyk, “Learning from imbalanced data: open challenges and future directions,” Prog Artif Intell, vol. 5, pp. 221-232, 2016.

[34] H. He, and E. A. Garcia, “Learning from imbalanced data,” IEEE Transaction on Knowledge and Data Engineering, vol. 21, no. 8, pp. 1263-1284, 2009.

[35] G. Menardi, and N. Torelli, “Training and assessing classification rules with imbalanced data,” Data Mining and Knowledge Discovery, vol. 28, no. 1, pp. 92-122, 2014.

[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” J Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[37] G. James, D. Witten, T. Hastie, and R. Tibshirani, "Tree-Based Methods," An Introduction to Statistical Learning: with Applications in R, pp. 303-336, New York: Springer, 2017.

[38] L. Breiman, “Random forests,” Machine Learning, vol. 45, pp. 5-32, 2001.

[39] M. Xiao, H. Yan, J. Song, Y. Yang, and X. Yang, “Sleep stage classification based on heart rate variability and random forest,” Biomedical Signal Processing and Control, vol. 8, pp. 624-633, 2013.

[40] L. Fraiwana, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, “Automated sleep stage identification system based on time-frequency Analysis of a Single EEG Channel and Random Forest Classifier,” Computer Methods and Programs in Biomedicine, vol. 108, pp. 10-19, 2012.

[41] Y. Bengio, and Y. Grandvalet, “No unbiased estimator of the variance of k-fold cross-validation,” Journal of Machine Learning Research, vol. 5, pp. 1089-1105, 2004.

[42] J.-H. Kim, “Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap,” Computational

Statistics & Data Analysis, vol. 53, no. 11, pp. 3735-3745, 2009.

[43] D. M. Powers, "Evaluation: from precision, recall and f-measures to ROC, informedness, markedness and correlation," J Mach Learn Technol vol. 2, no. 1, pp. 37-63, 2011.

[44] J. Hanley, and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Intell. Data Anal. J., vol. 143, pp. 29-36, 1982.

[45] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognition, vol. 30, no. 7, pp. 1145-1159, 1997.

[46] H. Guo, Y. Li, J. Shang, M. Gu, Y. Huang, and B. Gong, "Learning from class-imbalanced data: review of methods and applications," Expert Systems with Applications, vol. 73, pp. 220-239, 2017.

[47] P. Teetor, R Cookbook, p.^pp. 223: O'Reilly, 2011.

[48] C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft, and J. Popp, "Sample size planning for classification models," Analytica Chimica Acta, vol. 760, no. 14, pp. 25-33, 2013.

[49] Y. Benjamini, "Opening the Box of a Boxplot," The American Statistician, vol. 42, no. 4, pp. 257-262, 1988.

[50] O. Loyola-Gonzalez, J. Martinez-Trinidad, J. A. Carrasco-Ochoa, and M. Garcia-Borroto, "Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases," Neurocomputing, vol. 175, pp. 935-947, 2016.

[51] L. Zhou, "Performance of corporate bankruptcy prediction models on imbalanced dataset: the effect of sampling methods," Knowledge-Based Systems, vol. 41, pp. 16-25, 2013.

[52] A. Estabrooks, T. Jo, and N. Japkowicz, "Multiple resampling method for learning from imbalanced data sets," Computational Intelligence, vol. 20, no. 1, pp. 18-36, 2004.

[53] H. Guo, and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach," SIGKDD Explorations, vol. 6, no. 1, pp. 30-39, 2004.

[54] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," Pattern Recognition, vol. 48, pp. 1623-1637, 2015.

[55] J. Quinlan, "Improved estimates for the accuracy of small disjuncts," Mach. Learn., vol. 6, pp. 93-98, 1991.

[56] F. Li, X. Zhang, X. Zhang, C. Du, Y. Xu, and Y.-C. Tian, "Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets," Information Sciences, vol. 422, pp. 242-256, 2018.