NeonSleep: An open-source machine learning tool for estimating nightly sleep duration from smartphone usage log data

George Aalbers[1], Andrew T. Hendrickson[1], Mariek M. P. vanden Abeele[2], & Loes Keijsers[3]

[1] Department of Cognitive Science & Artificial Intelligence, Tilburg University

[2] imec-mict-UGent, Department of Communication Sciences, Ghent University

[3] Department of Psychology, Education & Child Studies/Clinical Child and Family Studies, Erasmus University Rotterdam

Author note

Correspondence concerning this article should be addressed to George Aalbers, Warandelaan 2, Tilburg. E-mail: h.j.g.aalbers@tilburguniversity.edu

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Permission number: REDC 2019.94b.

Conflict of Interest: George Aalbers, Andrew T. Hendrickson, Mariek M.P. vanden Abeele, and Loes Keijsers declare that they have no conflict of interest.

Acknowledgements: We wish to thank the Tilburg Experience Sampling Center (TESC) and the mobileDNA team at imec-mict-UGent for their assistance in setting up our digital phenotyping study. We thank our colleagues Ghaith Al Seirawan, Andrei Oprea, Ethel Pruss, Marieke van der Pol, and Kyle van Gaeveren for their outstanding help during data collection.

Funding: Funding was received from the Tilburg University IMPACT program as well as a VIDI grant (NWO VIDI grant no. 452.17.011) awarded to Loes Keijsers.

Abstract

Purpose. Smartphone use is closely aligned with the sleep-wake cycle. In the past decade, researchers have therefore built algorithms for passively tracking sleep using smartphone log data. The aim of our preregistered study was to replicate previous work and make our non-proprietary code, machine learning models, and data openly available. **Methods.** Participants (N = 165) completed a sleep diary for up to 60 days (6,590 total observations). In parallel, a dedicated smartphone application continuously logged their smartphone application usage. We trained machine learning models to estimate sleep duration (self-reported by participants) from smartphone activity in 15-minute windows between 21:00 and 13:00 (i.e., 105,440 hours of smartphone use). **Results.** We found moderately strong, positive correlations between model estimates of sleep duration and self-reported sleep duration in out-of-sample data (r_{median} = 0.63, median absolute error [MAE_{median}] = 40.04 minutes). This relationship differed from individual to individual, ranging from weakly negative to strongly positive (r = [-0.26, 0.86]). Rank-order correlations between sleep duration estimates of the most accurate model (random forest) and self-reported sleep duration exceeded r = 0.5 in 59.39% of individuals (MAE = [12.41, 113.39] minutes). **Conclusions.** Our study contributes by (1) replicating previous findings that sleep duration can be inferred from smartphone application usage log data, (2) highlighting how model accuracy may differ for different persons, and (3) providing openly available Python code, trained models, as well as (preprocessed, anonymized) data for further optimization of sleep estimation models (URL to Github repository).

Keywords: smartphone; sleep; machine learning; passive sensing; digital phenotyping; experience sampling

Introduction

Because sleep is vital for health [1], non-invasive technologies for continuously tracking sleep parameters (e.g., duration, interruptions, regularity) are highly meaningful for capturing clinically relevant information. Such passive sleep monitoring might help researchers answer questions about bidirectional relationships between sleep-related behaviors (e.g., sleep regularity) and health-related parameters, such as psychological well-being [2]. In practice, passive monitoring of sleep opens up possibilities to improve self-monitoring of, for instance, patients with mood, anxiety, and psychotic disorders [3]. This is an exciting development as it could 1) partially unburden patients from the daily task of active self-monitoring, 2) make sleep tracking more accurate compared to self-monitoring, and 3) be used for just-in-time interventions. For instance, in bipolar disorder, passively logged sleep could be used to diagnose, in real-time, whether a patient is entering a phase of (hypo)mania (characterized by first slightly, and later much increased activity), which would require an intervention as soon as possible.

As the human sleep-wake cycle closely aligns with smartphone application usage patterns [4], researchers in different fields leverage smartphone log data to estimate sleep-related variables, particularly chronotype [4], sleep onset, offset, and duration [5-16], and sleep quality [7, 17-18]. The present study builds on this scholarship by making three unique contributions. First, we provide the reader with a brief overview of methodological approaches taken in previous studies and describe how the field of sleep inference has evolved over the years. This is important to demonstrate which research practices are currently state-of-the-art. Second, we replicate evidence that smartphone app usage log data can be leveraged to estimate sleep duration, evaluating model accuracy on a person-by-person basis. Third, we provide openly available (a) non-proprietary Python code for preprocessing raw smartphone usage data and model training, (b) trained sleep inference machine learning models, and (c) preprocessed data (URL to Github repository). Because the utility of sleep inference models is to accurately infer sleep duration in individuals who did not participate in the present study, we take a subject-independent and subject-specific approach to model evaluation. Subject-independent evaluation means we train models on one subset of individuals and evaluate them on another subset. Subject-specific evaluation means we evaluate model accuracy on a person by person basis, resulting in accuracy metrics for each individual in our study.

Overview of Related work

To find the majority of relevant previous work, we searched Google Scholar as well as the university online library system using the following terms “sleep duration”, “sleep onset”, “sleep offset”, “sleep

inference", "smartphone", "smartphone log data", "machine learning", and "deep learning". We also consulted reference lists of articles we found or were already aware of. Our literature search shows researchers have used different approaches to estimating sleep from smartphone and wearable log data. The general pattern of suggests the following (see Table 1 for a complete overview):

1. **Population**: Research in this domain has mostly investigated convenience samples (i.e., students, academic staff). Limited work focused on clinical populations (e.g., schizophrenia).
2. **Sample size**: Sample sizes have grown substantially, from below 40 (early 2010s) to above 100 participants (up to $N = 198$; late 2010s-early 2020s).
3. **Measuring sleep**: Older studies used self-report surveys to measure sleep. More recent studies use multimodal assessment.
4. **Feature modality**: Earlier studies included larger numbers of feature modalities to estimate sleep, often leveraging different smartphone sensors (e.g., microphone). Recent studies either (1) focus on smartphone use or (2) mix smartphone use with physiological features.
5. **Models**: Older studies have applied rule-based algorithms, Bayesian models or machine learning models that do not model temporal dependencies of smartphone use. More recent studies use deep learning models that leverage temporal dependencies and have developed novel rule-based algorithms.
6. **Model evaluation:** Studies have taken two approaches to evaluate models: in-sample and out-of-sample. In-sample means that a model or algorithm is trained and tested on the same data. Out-of-sample evaluation differs from study to study and can be categorized as follows: (1) subject-independent (i.e., train model on individuals A - D and evaluate on individual E), (2) subject-dependent (i.e., train model on observations A - D and evaluate on observation E of each individual), or (3) both. Studies also differ in the metrics they use to evaluate model performance and whether they assess performance on a person-by-person basis.

Our study analyzes one of the largest samples thus far, using only non-invasive smartphone application usage data to estimate sleep duration. The added value of our preregistered study is that we evaluate models on out-of-sample data, on a person-by-person basis and make all (preprocessed) data, code, and trained models available. In this way, our work can be used in future research, either for improving the models we trained here or for using our models to extract sleep duration estimates from smartphone application usage log data.

Table 1. Sample and data characteristics, analytic methodology, and accuracies reported in previous sleep detection studies. Studies are sorted according to publication year, starting with the earliest year.

Study	Sample	Features	Sleep	Models (task)	Evaluation set	Target	Accuracy
[5]	Unknown (*N*=13, *t*=62)	Accelerometer	A	Rule-based algorithms (classification)	In-sample	State	*acc* > 0.87
[6]	Students (*n*=5, *t* = 7) Staff (*n=3, t* = 7*)*	App usage (time, duration) Light intensity Microphone Recharge events	SR	Non-negative least squares linear regression (regression)	5-fold cross-validation (CV)	Duration	± 42 minutes
[7]	Diverse (*N*=27, *t* = 31)	Accelerometer App usage Battery Light intensity Microphone Screen status Screen proximity	SR	Decision tree, Bayesian network (classification)	Leave-one-subject-out CV Leave-one-night-out CV	Duration Duration	± 64 minutes ± 49 minutes
[8]	Students (*N*=9, *t* = 28-93)	Screen status	SR	Rule-based algorithm (regression)	In-sample	Duration	< 45 minutes
[9]	Students (*n*=16) Staff (*n*=2) (*t* =19-100)	Accelerometer App usage Battery Day of week WiFi (location) Hour of day Light intensity Microphone Screen proximity Screen status	SR	Logistic regression, support vector machine, random forest (classification)	10-fold CV per individual	Duration	± 40.7 minutes
[10]	Diverse sample (*N*=208, *t* = 11-137)	In-phone activity Light intensity Location Motion Microphone	SR	Random forest (classification)	Stratified 10-fold CV (across individuals) 3-fold CV (per individual)	Duration	± 58 minutes
[11]	Unknown (*n* = 126, *t* = 14-28) Students	App usage (time)	A N	Bayesian model (classification)	In-sample	State	*acc* = 0.89

	(*n* = 324, *t* = 0)						
[12]	Patients (*N* = 18, *t* = 30)	Accelerometer	SR	Bayesian model (regression)	In-sample	Duration	*r* = 0.69
[13]	Students (*N* = 186, *t* = 30)	Accelerometer Actigraphy activity Light intensity Location Screen-on Skin conductance Skin temperature Calls (time) SMS (time)	A	Long short-term memory (LSTM) (classification)	Participant dependent (first 20% or last 20% of consecutive days as test set) Participant independent (80%/20%)	State Onset Offset	*acc* = 0.965 ± 5.0 minutes ± 5.5 minutes
[14]	Students (*N*=79, *t*=14)	App usage (time) Tappigraphy*	A SR	Linear regression (regression)	In-sample	Duration	R^2 = 0.36 (*r* = 0.6)
[15]	Students (*N*=18, *t*=107-233)	Audio usage Daily activities Daily calories Day of week Nap duration Part of a semester Previous night sleep duration Screen usage Sleep start time Temperature Time to fall asleep	A	General linear model (GLM) Generalized linear mixed model (GLLM) (regression)	Leave-one-subject-out CV Leave-one-night-out CV	Duration	*r* = 0.583
[16]	Students (*n* = 120) Staff (*n* = 78, *t* = 38)	App usage (time) Tappigraphy	A SR	Rule-based algorithm (regression)	In-sample	Onset Offset	*rho* = 0.82 ± 4 minutes
This study	Students (*N*=165, *t*=20-60)	App usage	SR	LASSO, SVR, RF, gradient boosting regression (GBR)	Nested 5-fold group CV	Duration	*rho* = 0.63 ± 40 minutes

Note: In column ***Sample***, "**n**" indicates the number of people and "**t**" indicates the number of sleep estimates per person. In column ***Sleep***, "**A**" stands for actigraphy-assessed sleep, "**SR**" for self-reported sleep, "**N**" for no sleep measures. In column ***Target***, "**State**" stands for a binary outcome variable with values "Asleep" and "Awake" with multiple time windows per day (e.g., is a person Asleep or Awake between 0:00 and 0:01),

"**Duration**" stands for a continuous outcome variable "Sleep duration" in minutes, "**Onset**" stands for a continuous outcome variable "Sleep onset" in minutes, "**Offset**" for a continuous outcome variable "Sleep offset" in minutes. In the column ***Accuracy***, "**r**" stands for Pearson's correlation coefficient, "**rho**" stands for Spearman rank-order correlation coefficient, and "**acc**" stands for the proportion of time windows that a model correctly identifies a person to be asleep or awake (e.g., is a person Asleep or Awake between 0:00 and 0:01).

* Tappigraphy represents a log of touchscreen events – i.e.,a timestamp of when a person touches their screen.

Materials and Methods

Participants

We follow reporting guidelines recommended for experience sampling studies [19]. For a preregistered data collection (Open Science Framework [OSF] [blinded URL], 23rd of March, 2019), we recruited 247 student participants, 165 of whom we included for analysis (59.39% female; median age = 21.17 years [*SD* = 2.94]). We excluded participants with operating systems other than Android on their primary phone and participants with less than 20 available data points.

Procedure

Ethical approval was issued by the [institution name blinded] Ethics Committee (approval code [blinded]). Participants were recruited through the university participant pool. After receiving online information through Qualtrics, having been offered a possibility to ask questions, and signing an informed consent form, participants followed online instructions to install two applications on their smartphone. After completing these instructions, participants attended an onboarding session in which we provided additional information, offered further opportunity to ask questions, and motivated participants to participate to the best of their ability.

Technology. All participants installed two applications on their Android device: Ethica Data and mobileDNA. Ethica Data is an application that prompts participants to complete brief surveys on their smartphone (i.e., *experience sampling*). MobileDNA is an application developed by imec-mict-UGent (mict.be) that unobtrusively logs a person's smartphone application use, smartphone notifications, and location (i.e., *passive logging*).

Sampling scheme. In a four-month period, Ethica notified participants five times a day for a maximum of 60 days (30 days in month one; 30 days in month four) at pseudo-random times between 8:30 and 10:30 to complete a brief sleep diary, while mobileDNA continuously logged smartphone application use. Following an initial push notification, each survey was available to the participant for 50 minutes. After 45 minutes, they received a reminder notification. After 50 minutes, the survey expired. Participants were allowed to catch up on one missed survey per day by starting and completing a new survey.

Monitoring Protocol. Following good practices in the field (see [19]), We actively monitored participant compliance and motivated participants with weekly emails containing personalized feedback. When a participant failed to complete several consecutive surveys, we sent an email to inquire why they could not comply with the study protocol and how any issues might be resolved. In a limited number of cases, participants

did not respond to such emails, in which case we contacted them through a phone call. Participants were compensated with course credits for research participation and were entered into a raffle comprising twenty prizes of 15 euro.

Compliance. Participants (N = 165) completed a total of 6,590 surveys, with a median study adherence of 42 surveys (70.00%; SD = 10.98), which is average for experience sampling studies [19]. Reasons for noncompliance ranged from technical difficulties (e.g., not receiving any notifications; broken or lost smartphone) to not being able to complete a survey (e.g., waking up too late; receiving a notification during work or lecture) to personal reasons (e.g., attrition due to COVID-19-related personal problems or collecting sufficient course credits).

Measures

Consensus Sleep Diary. We used three (slightly modified) items from the Consensus Sleep Diary (CSD; [20]) to measure participants' daily sleep duration. The first item requires a participant to report the time they got into bed (bedtime), the second item the time it took before they fell asleep (sleep latency), and the third to report the time they woke up (wake time). Sleep duration can be calculated by subtracting the sum of bedtime and sleep latency from wake time.

Analytic approach

We trained LASSO regression (LASSO), support vector regression (SVR), random forest regression (RF), and gradient boosting (GB) regression models to predict sleep duration (on day *t* in participant *i*) based on a given participant's smartphone application usage log data (on day *t-1* and day *t* in participant *i*). For instance, to estimate the sleep duration of the night from Monday (day *t-1*) to Tuesday (day *t*), we used smartphone application usage log data of Monday evening (day *t-1*) through Tuesday afternoon (day *t*).

Features

We extracted features from raw passively logged smartphone usage data that included the time an application was used and for how long, both measured in seconds. We downsampled raw smartphone application usage to a 15-minute timescale, calculating total time spent on the smartphone per 15-minute window for time windows between 21:00 and 13:00. We divided each observation by 900 seconds so that each value was within the range between 0 and 1. We analyzed a total of 105,440 hours of smartphone usage log data.

Cross-validation

For training and testing the four models, we applied nested cross-validation. We trained the model on 80% of participants, using five-fold randomized search cross-validation to tune model hyperparameters and identify one set of optimized parameters per model type (see Appendix Table 1 for optimized hyperparameters). Subsequently, we tested model performance on the remaining 20% participants. We followed this procedure until we had trained and tested the model on all participants. This means we trained and tested each model type five times.

Model evaluation

We evaluated model performance on hold-out test data by computing (1) Spearman rank-order correlations between self-reported and estimated sleep duration and (2) median absolute error (MAE) of estimated sleep duration. For each model type, we report aggregate statistics (one correlation across all out-of-sample observations in one cross-validation fold) as well as median and range of person-specific statistics (correlation across all out-of-sample observations for each individual).

Results

Descriptives

Participants reported an average of 7.84 hours of sleep per day (*SD* = 1.45). On average, participants reported going to bed at 00:53 (*SD* = 1.49), falling asleep 24.74 minutes later (*SD* = 29.37), and waking at 08:43 (*SD* = 1.22).

Across participants, hourly smartphone application use demonstrated a periodic pattern across the week (Figure 1, inspired by [4]), recurrently peaking between noon and midnight and reaching a minimum in the night interval. Although this periodicity is rather consistent throughout the week, pre-midnight smartphone usage appears more intensive on Sundays through Thursdays (i.e., nights before a weekday) compared to Fridays and Saturdays (i.e., nights before a weekend day). This daily cycle in smartphone use roughly matches the human wake-sleep cycle, suggesting sleep duration might possibly be inferrable from a person's smartphone application use across the day.

As Figure 2 illustrates, individuals differ in their average temporal patterns in smartphone application usage. Participant 1 uses their smartphone more or less consistently throughout the day, only sometimes using it after midnight. Participant 2 uses their smartphone more commonly after midnight and shows a mild decrease in smartphone usage across the day.

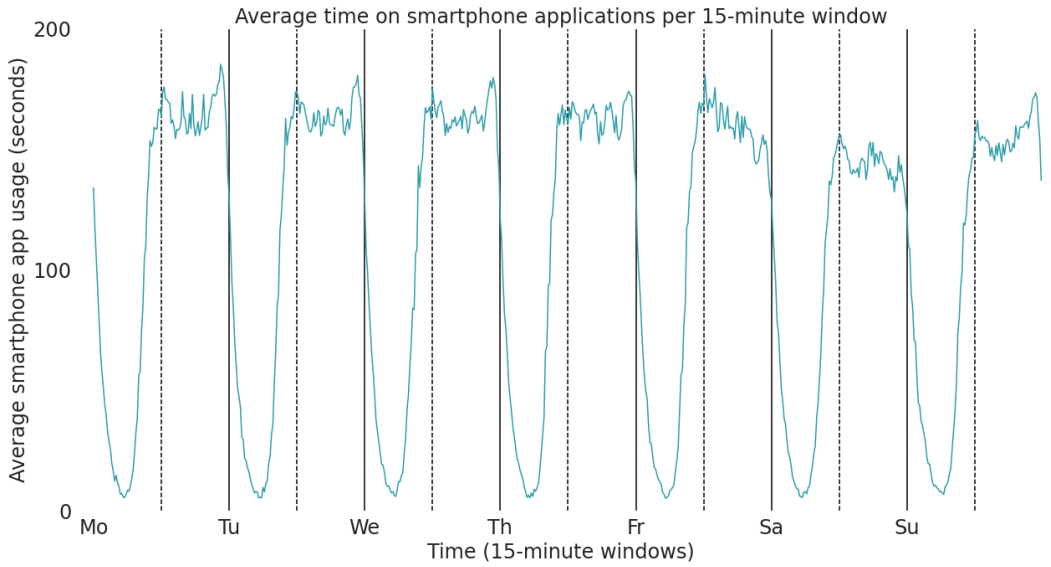

Figure 1. Mean duration spent on smartphone applications for each 15-minute time window of the week (i.e., 900 seconds), from Monday (left) to Sunday (right). The teal (solid) time-series represents smartphone usage per 15-minute time window (averaged across individuals), black solid vertical lines represent midnight, and black dotted vertical lines represent noon.

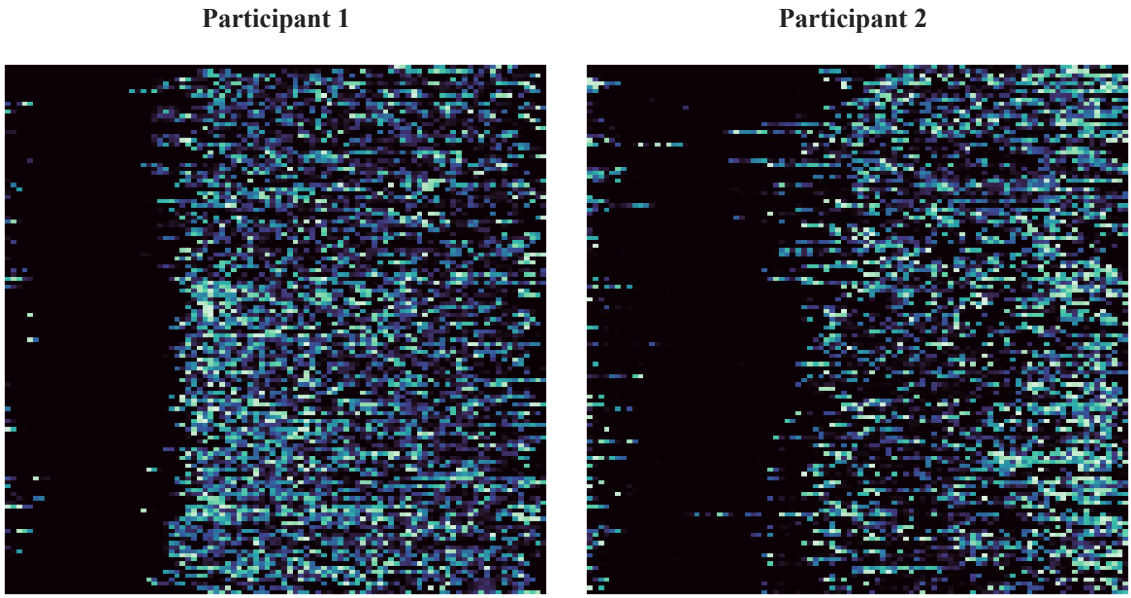

Figure 2. Visualizations of smartphone application usage duration per 15-minute time window for two participants (first 130 days per participant). In the heatmaps, each row contains data from one day and each column the usage during a 15-minute window (00:00 to 23:45). Lighter sections represent greater smartphone application usage duration, with black sections representing no usage.

Model performance

Overall, sleep duration estimates correlate positively with self-reported sleep duration for each model, with Spearman rank-order correlations ranging from $r_{overall} = 0.50$, $p < .001$ (LASSO) to $r_{overall} = 0.63$, $p < .001$ (RF). We report median and range for person-specific correlations per model in Table 2. Median person-specific correlations ranged from $r_{median} = 0.42$ (LASSO) to $r_{median} = 0.58$ (RF). Individual person-specific correlations ranged from $r_{min} = -0.26$ to $r_{max} = 0.86$ (both RF). Of these correlations in the RF, 9.09% (2.42% SVR; 2.42% GBR; 0.00% LASSO) were larger than 0.75, 59.39% were greater than 0.5 (40.61% SVR; 38.18% GBR; 36.97% LASSO), and 87.27% were greater than 0.25 (79.39% SVR; 75.76% GBR; 75.76% LASSO).

Table 2. Median and range of fold-specific and person-specific Spearman rank-order correlations between self-reported and model-based estimated sleep duration in out-of-sample data.

Model	Fold-specific *r*	Fold-specific *MAE*	Person-specific *r*	Person-specific *MAE*
LASSO	0.50 [0.33, 0.56]	44.66 [41.35, 46.62]	0.42 [-0.18, 0.81]	43.60 [12.86, 114.44]
SVR	0.52 [0.39, 0.59]	41.62 [39.25, 44.92]	0.46 [-0.14, 0.76]	41.39 [14.97, 131.96]
RF	0.63 [0.43, 0.67]	40.04 [39.77, 42.82]	0.58 [-0.26, 0.86]	40.32 [12.41, 113.39]
GBR	0.53 [0.39, 0.55]	42.28 [40.71, 43.38]	0.44 [-0.11, 0.80]	41.90 [21.68, 153.77]

Discussion

Principal findings

In a large longitudinal dataset (N = 165; 6,590 total observations), we investigated to what extent machine learning models can be trained to estimate self-reported sleep duration from passively logged smartphone use. We trained models on a subset of participants and tested their performance on another subset. Overall, model estimates correlated strongly positively with self-reports, but we found this relationship to vary from one participant to the next. Person-specific correlations were acceptable for our best model (random forest regression) in more than half of participants (rho > .5). The model's estimates of sleep duration tended to be off by approximately 40 minutes in the median participant. We make Python code for preprocessing the data, preprocessed data, and trained models openly available for future research (URL to Github repository).

Limitations

Our study has three important limitations. First, our ground truth (i.e., self-reported sleep duration) is widely accepted to be a noisy estimate of actual sleep duration ([13-14, 16]). This means our models might have learned to predict noise, despite extensive efforts to limit overfitting. Critically, such noise will also limit the association between ground truth and predicted values. That is, the theoretical maximum of accuracy should be lower than perfect for many participants, otherwise this indicates models were overfitted. It is thus possible trained models perform better than our results suggest. Second, we used an assumption-free approach to feature representation in that models were not aware features were temporally dependent. Third, time windowed features in this study might have been insufficiently granular. Possibly, models trained on minute-by-minute smartphone usage could be more successful [13]. In that case, it might be fruitful to train convolutional neural networks (CNN) or long short-term memory (LSTM) recurrent neural networks (RNN) to detect whether a person is asleep or awake and derive sleep duration from these predicted values (e.g., [13]).

Comparison with prior work

Compared to previous machine learning work in this domain, our model was trained on relatively low-cost, non-invasive passively logged data sources, using a large dataset. Although it is challenging to make a perfect one-on-one comparison with previously presented models, our results suggest our model is on par with and might outperform models with access to more information, such as microphone and light intensity [7, 10]. This is of interest as logging only application usage is less invasive, requires less memory storage, and is less battery intensive than logging multiple sensors [11]. Notwithstanding this improvement, our best model does not outperform all previous efforts in the domain. First, it is likely less accurate than a long short-term memory

recurrent neural network (LSTM RNN; [13]) that was trained on multiple data sources (e.g., actigraphy, location, skin temperature, screen status) and detects whether a person is awake or asleep with nigh-perfect accuracy (96.5%). The multimodal nature of these data makes this state-of-the-art model expensive to deploy, making it more suitable for applications in a clinical population (e.g., bipolar disorder) than in the general population, for which our model might be more suitable. Second, our model is less accurate than a recently proposed rule-based algorithm based on screen status and tappigraphy data (i.e., a log of when a person touches their screen; [16]), likely because this algorithm has access to more granular data. A disadvantage of this algorithm is that it can only be used after a person's smartphone usage has been logged for a while, as it utilizes an estimate of a person's most likely period of sleep for selecting sleep onset and offset. By contrast, our model does not require any person-specific training data and can immediately be used to estimate sleep duration for new people.

Conclusions

The main conclusion from this study is that supervised machine learning models can estimate sleep duration based on smartphone usage patterns, providing further support to previous claims that smartphones might be used to monitor sleep. Using a relatively non-invasive, affordable, and scalable data source, our best model outperforms many previously trained machine learning models. The approach taken in this study can be extended to include other data sources and algorithms and might be viewed as a step towards building an open platform that brings together rich behavioral data, non-proprietary code, and models to infer sleep from digital traces.

References

1. Jha, VM, Jha SK. Sleep Loss: What Does It Do to Our Brain and Body?. In: Sleep: Evolution and Functions. Springer: Singapore; 2020. pp. 61-78.
2. Fischer D, McHill AW, Sano A, et al. Irregular sleep and event schedules are associated with poorer self-reported well-being in US college students. Sleep. 2020; 43:zsz300.
3. Aledavood T, Torous J, Triana Hoyos AM, et al. Smartphone-based tracking of sleep in depression, anxiety, and psychotic disorders. Current psychiatry reports. 2019; 21:1-9.
4. Aledavood T, Kivimäki I, Lehmann S, et al. A non-negative matrix factorization based method for quantifying rhythms of activity and sleep and chronotypes using mobile phone data. arXiv preprint arXiv:2009.09914. 2020.
5. Natale V, Drejak M, Erbacci A, et al. Monitoring sleep with a smartphone accelerometer. Sleep and Biological Rhythms. 2012; 10:287-92.
6. Chen Z, Lin M, Chen F, et al. Unobtrusive sleep monitoring using Smartphones. 7th International ICST Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health). 2013: .
7. Min JK, Doryab A, Wiese J, et al. Toss'n'turn: smartphone as sleep and sleep quality detector. InProceedings of the SIGCHI conference on human factors in computing systems. 2014: 477-486.
8. Abdullah S, Matthews M, Murnane EL, et al. Towards circadian computing: "early to bed and early to rise" makes some of us unhealthy and sleep deprived. Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing. 2014: 673-684.
9. Huang K, Ding X, Xu J, et al. Monitoring sleep and detecting irregular nights through unconstrained smartphone sensing. IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing. 2015: 10; 36-45.
10. Saeb S, Cybulski TR, Schueller SM, et al. Scalable passive sleep monitoring using mobile phones: opportunities and obstacles. Journal of medical Internet research. 2017; 19:e6821.
11. Cuttone A, Bækgaard P, Sekara V, et al. Sensiblesleep: A bayesian model for learning sleep patterns from smartphone events. PloS one. 2017; 12:e0169901.
12. Staples P, Torous J, Barnett I, et al. A comparison of passive and active estimates of sleep in a cohort with schizophrenia. NPJ schizophrenia. 2017; 3:1-6.
13. Sano A, Chen W, Lopez-Martinez D, et al. Multimodal ambulatory sleep detection using LSTM recurrent neural networks. IEEE journal of biomedical and health informatics. 2018; 23:1607-1617.

14. Borger JN, Huber R, Ghosh A. Capturing sleep–wake cycles by using day-to-day smartphone touchscreen interactions. NPJ digital medicine. 2019; 29:1-8.
15. Chen CY, Vhaduri S, Poellabauer C. Estimating sleep duration from temporal factors, daily activities, and smartphone use. In2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC). 2020: 545-554.
16. Massar SA, Chua XY, Soon CS, et al. Trait-like nocturnal sleep behavior identified by combining wearable, phone-use, and self-report data. NPJ digital medicine. 2021; 4:1-10.
17. Sano A, Phillips AJ, Amy ZY, et al. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN). 2015: 1-6.
18. Sathyanarayana A, Joty S, Fernandez-Luque L, et al. Sleep quality prediction from wearable data using deep learning. JMIR mHealth and uHealth. 2016; 4:e6562.
19. van Roekel E, Keijsers L, Chung JM. A review of current ambulatory assessment studies in adolescent samples and practical recommendations. Journal of Research on Adolescence. 2019; 29:560-77.
20. Carney CE, Buysse DJ, Ancoli-Israel S, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. Sleep. 2012; 35:287-302.

Appendix

Table 1. Overview of hyperparameters per model

Model	Hyperparameters	Range
LASSO	Alpha	[0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 10000]
SVR (radial basis function kernel)	Gamma	[“scale”, “auto”]
	C	[0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000]
	epsilon	[0.0001, 0.001, 0.01, 0.1, 0.3, 0.5, 0.7, 0.9]
RF	Number of estimators	[500, 1000, 2000]
	Max features	[2, 7]
	Max leaf nodes	[2, 5, 10]
	Max depth	[1, 5, 10]
	Min samples split	[0.001, 0.01]
	Min samples leaf	[0.001, 0.01]
	Min weight fraction leaf	[0.001, 0.01]
GBR	Learning rate	[0.1, 0.2, 0.3, 0.4]
	Alpha	[0.1, 0.2, 0.3, 0.4]
	Number of estimators	range(800, 1000, 2)
	Max features	[“auto”]
	Max depth	[60, 62, 65, 67, 70]
	Min samples split	[10, 50, 100]
	Min samples leaf	[2, 4]

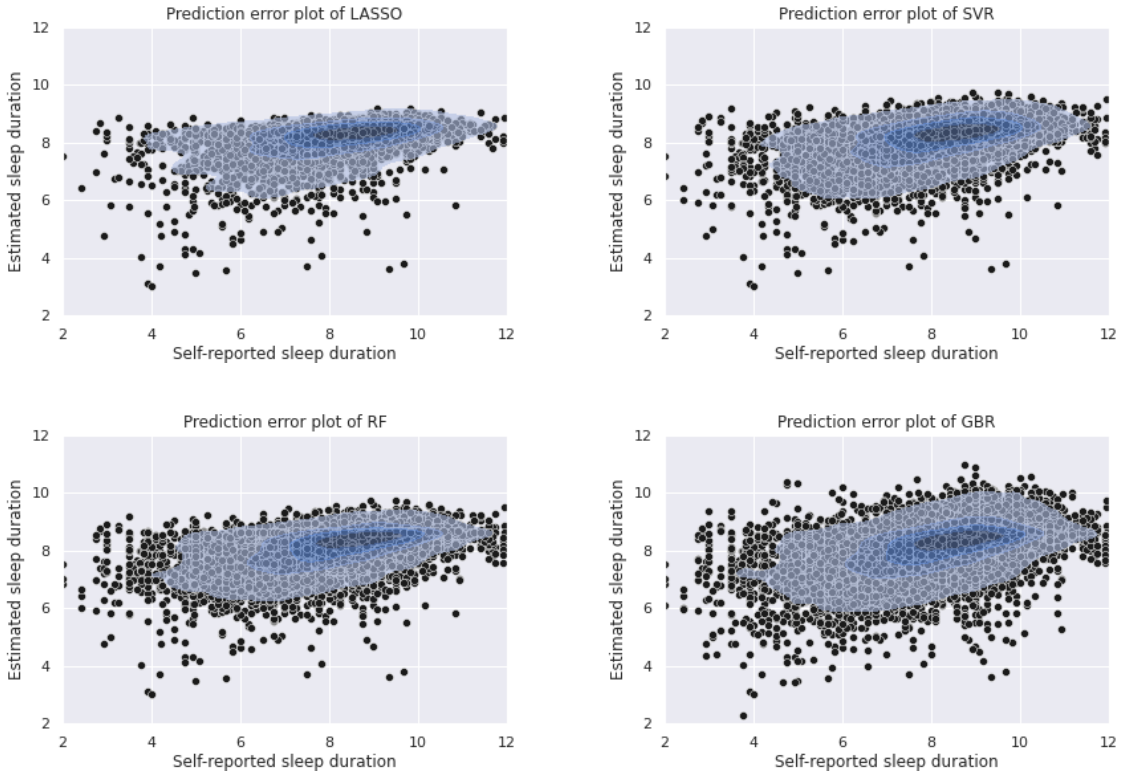

Figure 1. Prediction error plot for estimated sleep durations (y-axis) versus self-reported sleep duration in hours (x-axis) per model. Darker areas represent areas with greater density – most self-reported and estimated sleep durations are in this area.

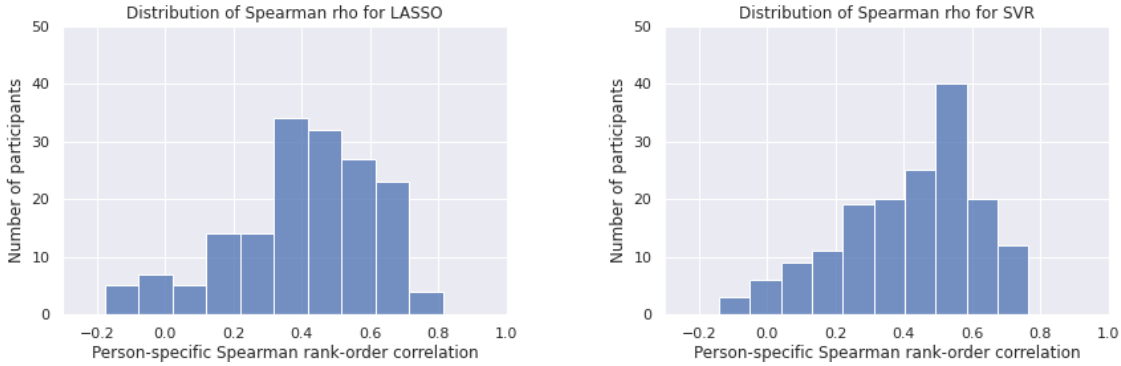

Figure 2. Strength of association between self-reported and estimated sleep duration in out-of-sample individuals. These histograms demonstrate to what extent model estimates of sleep duration would be concordant with self-reports of sleep duration in new settings and samples. Positive associations with greater strength represent stronger concordance for one individual.

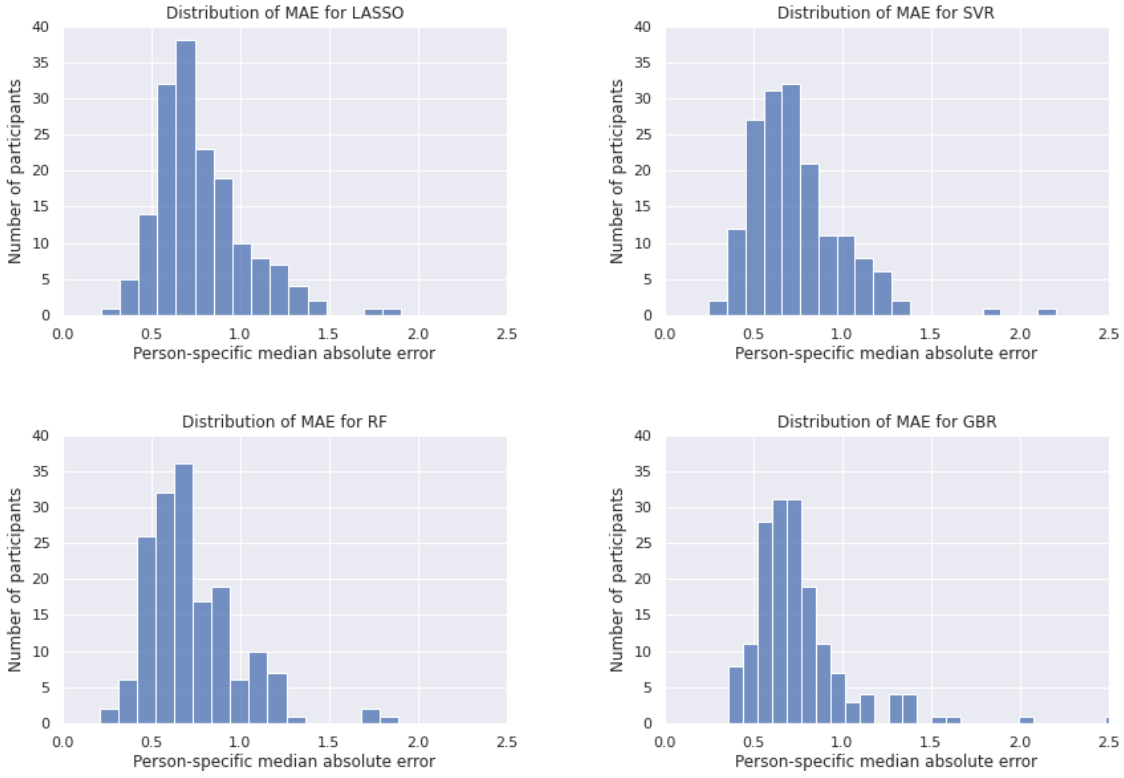

Figure 3. Accuracy of model estimates per individual. These histograms show how accurately models would estimate sleep in new settings and samples, measuring the discrepancy between estimated and self-reported sleep duration as the mean absolute error (MAE) in hours. Smaller values indicate that a model estimates (self-reported) sleep duration more accurately for a given individual.