SleepMore: Inferring Sleep Duration at Scale with Multi-Device WiFi Sensing

CAMELLIA ZAKARIA, University of Massachusetts Amherst, USA
GIZEM YILMAZ, National University of Singapore, Singapore
PRIYANKA MAMMEN, University of Massachusetts Amherst, USA
MICHAEL CHEE, National University of Singapore, Singapore
PRASHANT SHENOY, University of Massachusetts Amherst, USA
RAJESH BALAN, Singapore Management University, Singapore

Sleep provides the foundation for our daily routines. Despite the broad array of wearable and contactless sensing devices to monitor sleep, tracking sleep habits continuously over long periods remains challenging. In this paper, we leverage ubiquitous mobile devices and passive WiFi sensing to address the limitations of current approaches. Since modern Internet users own multiple mobile devices, we present an approach that leverages passive observations of WiFi activity from multiple mobile devices owned by a user to predict their sleep duration accurately. *SleepMore* provides an accurate, easy-to-deploy sleep tracking approach based on machine learning over the user's WiFi network activity. It first employs a semi-personalized random forest model with an infinitesimal jackknife variance estimation method to classify a user's network activity behavior into sleep and awake states at a fine-grain minute granularity. The system then uses these state sequences to estimate, using a moving average, the user's nocturnal sleep period and its uncertainty rate. Uncertainty quantification enables *SleepMore* to overcome the impact of noisy WiFi data that can yield large prediction errors. We validate *SleepMore* using data from a month-long user study involving 46 college students and compare it with an Oura Ring wearable. To demonstrate the applicability of our approach beyond college campuses, we also evaluate *SleepMore* on a private homeowner for over a month. Our results demonstrate *SleepMore* produces statistically indistinguishable sleep statistics from the Oura ring baseline. Predictions made within a 5% uncertainty rate range between 15-28 minutes of sleep error and 7-29 minutes of wake error, proving statistically significant improvements over prior work. We also conduct an in-depth analysis to identify the sources of errors.

1 INTRODUCTION

Sleep is essential to one's physical, emotional, and mental health. The National Sleep Foundation (NSF) recommends between 7 to 9 hours as appropriate for adults to maintain general wellness [19]. However, sleep deficiency is a common public health problem, with approximately 50 to 70 million Americans suffering from chronic sleep disorders [2]. The risk factors associated with insufficient sleep are performance and cognitive deficits, and its long-term effects are correlated with severe consequences such as obesity, stress, depression, and stroke [2, 19]. Thus, accurately determining sleep habits remains a topic of immense interest.

The gold standard for measuring sleep is using polysomnography (PSG), a multichannel, multimodal approach performed in controlled environments like a sleep clinic by trained technicians. These factors make PSG challenging to use for day-to-day sleep monitoring. More recently, wearable devices (e.g., FitBit, Jawbone), which are user-friendly and low-cost sleep, [7, 11, 45] have become popular for sleep monitoring. However, wearable devices require behavior changes by requiring users to wear the device when sleeping, which many users are reluctant to do [11]. Users who have the most disrupted sleep experiences tend to be those who do not wear sleep trackers. To overcome the challenges of continuous sleep monitoring on a longitudinal basis, researchers have explored contactless methods using radio, and radar signals [20, 44]. However, the requirements

Authors' addresses: Camellia Zakaria, University of Massachusetts Amherst, USA, nurcamellia@cs.umass.edu; Gizem Yilmaz, National University of Singapore, Singapore, gizem.yilmaz@nus.edu.sg; Priyanka Mammen, University of Massachusetts Amherst, USA, pmammen@cs.umass.edu; Michael Chee, National University of Singapore, Singapore, michael.chee@nus.edu.sg; Prashant Shenoy, University of Massachusetts Amherst, USA, shenoy@cs.umass.edu; Rajesh Balan, Singapore Management University, Singapore, rajesh@smu.edu.sg.

to instrument building infrastructure limit their scalability. Recently, using WiFi signals such as channel state (CSI) and backscatter information has been proposed as viable solutions for sleep tracking [54]. This is made possible by both smartphones becoming common across all income levels[39] and WiFi solutions becoming prevalent in public, institutional and residential locations [36, 40]. These solutions attempt to detect when the user has fallen asleep by checking for vibrations and other signals. Unfortunately, even though these solutions have reasonable accuracy, they require custom hardware and are thus hard to deploy at scale.

Smartphones have been used for sleep detection and monitoring. Prior approaches use smartphone activity as a proxy for user activity and long periods of device inactivity to infer sleep periods. These approaches include both client-side methods, based on monitoring screen activity [8], as well as network-side methods, based on monitoring WiFi network activity of the device [26]. Since users tend to habitually use their mobile device before they sleep and upon waking up, device activity, or lack thereof, is a feasible approach for detecting sleep periods [49]. While prior work has shown the feasibility of such approaches, they are known to suffer from significant errors, often more than an hour, when determining sleep time. These challenges motivate *the need to develop more accurate and generally accessible sensing approaches that monitor users' daily sleeping behavior without changing their routines*. In addition, such functionality could also contribute as a critical resource to the longitudinal requirement in almost all sleep medicine studies, facilitating precise data collection of fundamental sleep measures in real-time.

In this paper, we present *SleepMore*, a practical approach to sleep monitoring using passive observations of a user's device WiFi activity. Our approach leverages the growing number of mobile devices owned by each user in recent years (e.g., phone, tablet, laptop, e-readers) [9]. While smartphones remain the primary mobile device for most users, users may use a combination of devices over the day (e.g., use a tablet to stream content prior to bedtime and use the phone for other activities). We hypothesize that the higher errors of a single device sleep detection approach [8, 26] can be overcome by observing network activity from all the user's devices to infer sleep and wake times accurately.

SleepMore collects network activity information of all devices directly from the WiFi access points (AP) as features. Then, it employs a two-pronged technique, running both a random forest machine learning classification and moving average estimation models to predict sleep duration. Specifically, it first classifies users in *sleep* or *awake* states and computes confidence intervals for these predictions using an infinitesimal jackknife variance estimation method. Outcomes with less than 95% confidence level are noted as low confidence predictions. From applying a moving average, the most extended sequence of sleep states is observed as the user's nocturnal sleep period, with the start of the sequence denoting bedtime, T_{sleep}, and the end of sequence denoting wake time, T_{wake}. Simultaneously, the uncertainty rate of this estimation is instanced by the number of low confidence prediction states present in this sequence.

Through an IRB-approved study conducted over four weeks during an academic semester, we evaluated the performance of our system among 46 undergraduate students who resided on campus and one private homeowner. It is essential to underscore that our study defined no inclusion/exclusion criteria, especially on their sleep habits. That is to say, participants were not required to report receiving a formal sleep disorder diagnosis (e.g., sleep apnea) and were a mix of habitual and irregular sleepers. As part of the study protocols, the student participants wore the *Oura ring* (gen 2) [33] wearable sleep tracker for baseline. They were required to connect to the campus WiFi while in their residential dormitories. Participants supplied the WiFi MAC addresses of the multiple devices they owned (i.e., smartphones, tablets and laptops) so that we could identify these devices directly from the WiFi infrastructure and extract their network event logs (by default, all WiFi traces are anonymized). Our homeowner participant provided their WiFi network logs. Thus it achieves scalability and accuracy with no active user participation and only requires that users connect their devices to the same WiFi network. To the best of our knowledge, this is the first work to accurately predict sleep using a scalable WiFi-based technique using inputs from multiple user-owned devices.

In designing, implementing, and evaluating *SleepMore*, our paper makes the following contributions:

(1) We present a random forest machine learning-based algorithm with infinitesimal jackknife variance estimation and moving average smoothing technique to predict users' nocturnal sleep from WiFi network activity data of multiple user devices in residential spaces. Our ML classifier can predict the state of a user as sleep or awake and estimate sleep duration based on the most extended sequence of sleep states within a 24-hour period. We employ the variance estimation method to measure how confident a prediction is being made, flagging predictions beyond a 95% confidence interval as low-confidence outcomes and calculating the uncertainty rate of sleep estimates by the number of low-confidence outcomes present in the sequence of sleep states. These techniques accurately estimate a user's sleep duration and determine their bedtime and wake time.
(2) We implement *SleepMore* as a cloud-based web service, building a semi-personalized model that requires 40% of users' data for training, equivalent to 9 days of training data for 23 days of prediction. We deploy our solution in two residential settings — student dormitories and private homes, thus, demonstrating the feasibility of our approach in real-world settings and actual residents.
(3) We conduct an extensive experimental evaluation of *SleepMore* on campus and in a home. With humans sleeping on average 20% of the time in a day, our results show that *SleepMore* can accurately determine the state of users sleeping with approximately 90% recall. These state predictions help form accurate estimations of users' sleep duration and bed and wake times. We characterize key factors that impact the performance of our approach, including the use of one to many devices and the lack of training data to learn night-owl sleep schedules that did not often occur among our participants.
(4) We demonstrate *SleepMore* estimating nocturnal sleep duration, achieving statistically indistinguishable ($p>.1$) statistics compared to the state-of-the-art sleep tracker, Oura ring. Predictions made within a 5% uncertainty rate range between 15-28 minutes of sleep error and 7-29 minutes of wake error, proving statistically significant improvements over prior work. Note that predictions within 5% uncertainty rate make up more than 80% of our predicted outcomes. It is also essential to note that the reported accuracy obtained is similar to that of a baseline dedicated sleep device and not as ground truth for medically approved sleep studies requiring highly comprehensive and expensive polysomnography. Given its comparable performance to the Oura ring, *SleepMore* offers a promising complementary tool in supporting consumers' everyday sleep behavior and longitudinal studies for sleep medicine research without requiring the issuance of wearables or interfering with their natural behavior.

2 MOTIVATION AND BACKGROUND

This section provides background on sleep monitoring applications and sensing techniques, and motivates our multi-device passive-sensing approach.

2.1 Sleep Health, a Call-for-Action

Much work has reported sleep deprivation as a public health burden [2, 35]. Researchers sought to understand the reasons and consequences of insufficient sleep in different populations by measuring standard sleep parameters such as time in bed, sleep duration, wake frequencies, and sleep latency [2, 24]. These studies have noted insufficient sleep among adults as a result of lifestyles and work schedules [2], while adolescents' sleep loss is positively associated with more device use and online activities [46]. At the very least, *sleep duration is documented as the most fundamental and critical predictor of different health outcomes* with longitudinal associations to weight gains [25], quality of life [32], cardiovascular illnesses [5], cognitive impairments [14], to name a few.

The paradigm shift of recognizing sleep as a critical predictor of significant health consequences [16] has led to a fast-growing trend of consumer products offering digital sleep health options for monitoring and improving sleep.

From a research perspective, it has spurred a clear call for action among clinicians to assess sleep health for various age groups comprehensively. Many of these works raised concerns over support for the basic understanding of sleep, specifically, improving the effectiveness of sleep screening [2, 35] over longitudinal periods and developing new technologies to accomplish this task [30].

2.2 Sleep Sensing Technologies

In clinical practice, sleep is generally evaluated using validated retrospective scales, daily sleep logs, and/or polysomnography. However, it is challenging for everyday consumers to personally monitor their sleep behavior with these methods. Further, it remains highly burdensome for sleep studies to conduct long-term evaluation. Hence, sleep technologies offer a low-burden approach to automatically detect a user's sleep in the comfort of their own homes.

Contact-based wearables Commercial alternatives have shown feasibility in monitoring sleep. Zambotti's defined *sleep wearable trackers* as "over-the-counter, relatively low-cost devices available without prescription or clinical recommendations," [11] varying from wristbands to smartwatches, earbuds to rings. The study protocol employing sleep wearables typically involves loaning these devices to enrolled participants over a fixed duration for reusability in future studies. The implications of such practice could result in behavior modification during the study, from users not being used to wearing a device to bed and raising the challenge of a high user attrition rate. Over long monitoring periods, researchers reported the need for regular "use the wearable" reminders [11, 27, 56]. Further, limited data acquisition on sleep behavior is attributable to returning the device (e.g., after 15 days [6]) and not retaining a more standard modality option to persist the sensing operation.

Contactless sensing Low compliance with wearables has motivated the design of contactless sensing approaches. Examples include the use of cameras [21], RF or radar sensing to characterize sleep stages [44] and posture [55]. Despite these advances, RF sensing systems have yet to accelerate commercial adoption, while camera-based solutions are more likely to intrude on privacy interests.

Smartphone-based sensing By contrast, smartphone-based sensing had arose as an option from the developing smartphone dependency [37] among everyday users, leading to researchers using the "phone-as-a-sensor". Efforts specific to sleep monitoring include distinguishing respiratory patterns through a microphone [48] or inferring user sleep behavior from monitoring screen interactions and application usage [1, 10, 15, 17, 29]. In all these cases, the solution presents a dedicated mobile application that must be installed in users' smartphones and its data acquisition running in the device's background. Due to continuous data collection affecting device performance and privacy concerns, studies have also reported higher chances of user dropout over long periods [50].

Passive WiFi-based sensing Attempts to bypass the high attrition rate in smartphone-based sensing led to the utilization of WiFi-based passive sensing techniques. Prior work by Mammen *et al.* has shown that it is possible to use WiFi connection logs through users' single smartphone collected from the WiFi infrastructure to predict sleep [26], however yielding only 88.50% accuracy. Separately, broader surveys on behavioral monitoring via smart devices have argued that utilizing single-device for monitoring is not fully comprehensive, in part because of inadequate device coverage from users owning multiple devices [37, 38, 51]. Our work aims to extend prior work in the following ways:

(1) Prior work only uses data collected from a single smartphone based on the assumption that users spend most of their time online and on their smartphones [52]. We hypothesize that **including all devices owned by the user will significantly improve the accuracy of sleep detection**.

(2) This is a significant gap between the accuracy of smartphone-based methods and those achieved by sleep wearable trackers such as the Oura Ring which can achieve 96% recall at detecting sleep [3] With this result in mind, we aim to propose a solution that also uses WiFi sensing but **can significantly outperform current prediction standard and achieve no statistically significant difference with a wearable sleep tracker**. In this study, we use the Oura Ring (gen 2) as a representative wearable device for sleep monitoring.

2.3 Design Rationale

Figure 1 shows an example of the network connection frequency for a typical user with multiple devices every 15-minutes through 24 hours between Day_1, 6 pm to Day_2, 6 pm. The observation that the personal smartphone is the last device used before bedtime [49] makes it viable as a sleep monitoring sensor. However, users tend to own multiple devices, and in this case, a secondary tablet device denotes the first active usage upon the user waking up. This behavior presents practical reasoning to infer users' sleep behavior more accurately through multiple device usage. Hence, this work utilizes the WiFi network device activity collected from multiple devices to estimate an individual's nocturnal sleep duration over a longitudinal basis as the most fundamental feature supporting sleep studies.

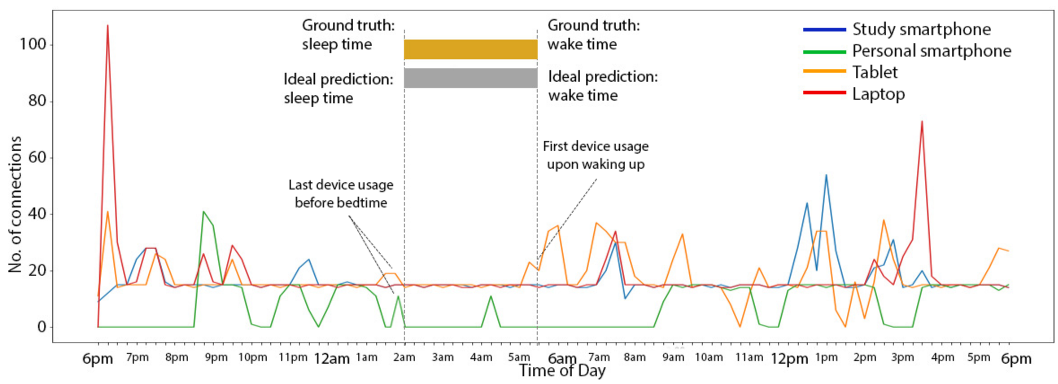

Fig. 1. Utility of smartphone and other devices before sleep and upon waking up. Low WiFi device activity from multiple devices corresponds to a sleep behavior.

Specifically, it monitors the devices that are associated with the network to infer the user's bedtime and wake-up time (referred to as T_{sleep} and T_{wake}). Adding new user-owned devices is easy as they would only need to be connected to the same WiFi network. In addition, our solution significantly preserves privacy as it does not require decoding the actual content of any packets sent by a device. For each user-owned device (denoted by their MAC address), we capture the WiFi connection events that these devices make to the network and compute the frequency at which each device establishes network connections.

The above example yields several critical insights: (1) The connection frequency increases with active online device utilization and decreases with less utility. For example, we can observe the user displaying the highest device use between 6-7 pm using a tablet. (2) A user also switches between four devices throughout the 24 hours, highlighting the potential for a more comprehensive behavioral monitoring by expanding device selection. (3) The user's smartphone is the last device used before bedtime, but it is not always the first device used upon waking up. (4) Further, network activity for other devices such as the laptop is observed to pick up soon after. (5)

While the frequency of connection increases with more device use, it is important to note that increased device activity does not always imply a user being active in true nature. For example, it can occur from an application on the device running automatic software updates or accepting notifications (e.g., incoming emails, received online messages), as demonstrated by the user's smartphone network connection log, which peaked between 4 to 5 am. These observations suggest that multi-device monitoring will lead to significantly better sleep prediction than using a single device. Indeed, our results in Section 5 confirm this.

It is important to emphasize that our approach aims to infer users' sleep duration by estimating their sleep and wake times, achieving robustness across users. The properties of utilizing a coarse-grained data source will innately restrict our technique in predicting more nuanced sleep characteristics such as the REM stages of sleep. Hence, our technique aims not to replace existing sleep-sensing modalities that offer fine-grained information but instead complement the use of such modalities in supporting sleep monitoring, especially over long periods.

2.4 Key Systems Challenges

Our proposed approach demonstrates non-trivial challenges in using device network events as a sleep predictor, as shown in Figure 2.

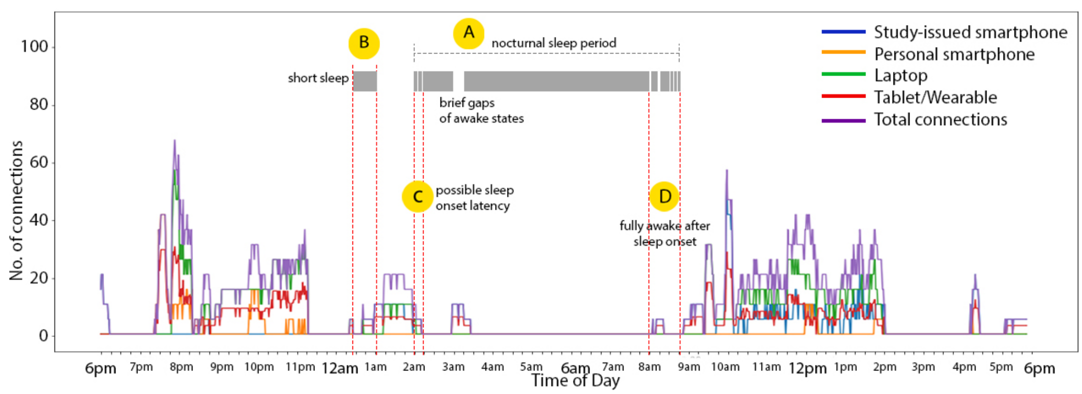

Fig. 2. Challenges predicting sleep with WiFi data.

2.4.1 *Noisy Data.*

Prior work has discussed how noise in WiFi-sensing, such as ping-ponging of AP connections, can negatively impact prediction performance [26]. In our case, the AP ping-pong effect is largely minimized by limiting WiFi network data collection to within users' residential premises. However, our system will still be susceptible to other noise-affected errors. Specifically, our technique assumes that increased WiFi network activity on a user's device indicates the user is awake. It is also reasonable to assume a user interacting with different devices before bedtime and upon waking up [13, 22]. However, this increased activity can be caused by device updates and app notifications that do not reflect the user's actual physical state. For example, the user is asleep at 3:00 am while the device updates or apps receive notifications (A). The variability of WiFi network activities resulting from both user and device actions will, to a large extent, affect our technique's ability to accurately predict a user's sleep duration.

Our system is designed to overcome these challenges by estimating the confidence intervals for predicting sleep states throughout a 24-hour period. While sleep states with less than 95% confidence level are noted as low confidence predictions, it continues to estimate the user's nocturnal sleep duration based on the longest sleep state sequence. However, it calculates the uncertainty of this estimation based on the number of low confidence predicted states present in this sequence. We provide the details to our technique in Section 3.3.1.

2.4.2 Accurate Estimate of Bed and Wake Times. While it is not unusual for a person to use their device before bedtime, it is also not unusual for a person to take some time to fall asleep after putting their device away [13, 22]. This duration is known as *sleep onset latency* and is estimated to be within 20 minutes. However, it also increases progressively with age and electronic device-use [4, 18, 23]. In Figure 2, the user is predicted to sleep briefly between midnight and 1:00 am before going back to bed at 2:00 am (A and B). Thus, the challenge is determining the true start of a user's bedtime. Further, the first predicted sleep state may not reflect that the user is falling asleep. Conversely, the last predicted sleep state may not imply a user waking up (C and D).

This challenge informs our decision to estimate the user's nocturnal sleep duration based on the most extended sequence of sleep states, as mentioned above. Detailed in Section 3.3.2, we employ moving average to smooth out short-term occurrences in wake states and highlight longer-term sleep trends over the 24-hour period.

3 *SLEEPMORE* DESIGN

This section presents our design of *SleepMore* as a multi-device WiFi sensing approach for sleep monitoring. Figure 3 presents the system overview. We describe the implementation details as follows.

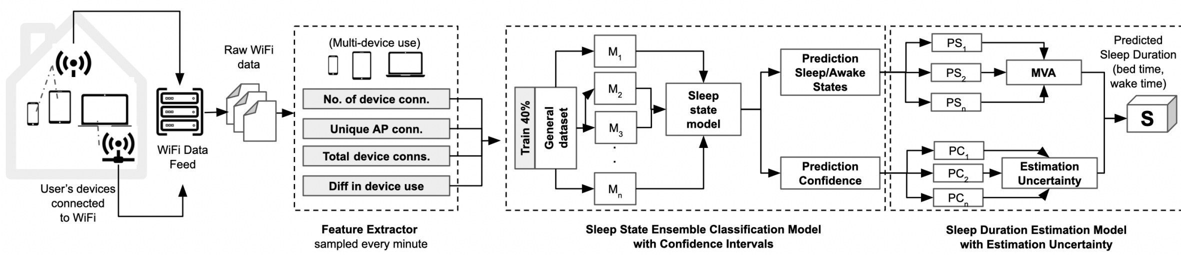

Fig. 3. *SleepMore*, a two-step process in predicting sleep states with confidence interval, and estimating bed and wake times.

3.1 Collecting WiFi Data Using RTLS

SleepMore requires WiFi data indicating when a user's device(s) communicate on the network. To obtain this in a scalable way, we leverage the real-time location system (RTLS) data feed available on commercial access points. For this paper, we collected all our data from a university dormitory that uses Aruba WiFi APs that support RTLS [31]. The RTLS feed provides reports from every AP that sees a device communicating on the network. Each RTLS message contains the following information:

```
Timestamp, Packet Age, Data Rate, DeviceType, Channel, DeviceMACAddress,
AssociationStatus, RSSI, Noise Floor, BSSID, MON BSSID
```

SleepMore only uses data from associated client devices as those clients have legitimate access to the network. In particular, we discard all data from unassociated devices as they are primarily WiFi probe requests from clients roaming for a usable network. Note: while RTLS feeds can be used to track the location of WiFi devices, *SleepMore* is only using this data as a proxy for device activity.

The RTLS reports are generated by all APs every five seconds in our test environment. However, these reports will only list the WiFi devices that were active and seen by that AP (i.e., they were using the same WiFi frequency as the AP and close enough to that AP) in that five-second interval. In particular, if a device has gone to sleep (e.g., because the user is sleeping and the device is charging), the device will not be emitting any network packets. It will not be seen and reported by any RTLS report generated by any AP.

3.2 Pre-processing Module

The first step in *SleepMore*'s pre-processing pipeline is to clean the noisy WiFi RTLS data. In particular, we remove data with invalid timestamps or RSSI values too weak (indicating spurious transmission). We only retain records of multiple device WiFi activities on days that users provide their Oura data (see Table 1 for data summary).

SleepMore's features are generated from the RTLS *Timestamp* (records time at which a data message is received), *Device Type* (client or AP station), *Device MAC Address* (client or AP station), and *Association status* (associated or unassociated device) fields. Because a user could own multiple devices, each with their own MAC address, *SleepMore* generates device features primarily based on the number of associated connection events per device (i.e., *networkEvents*) and the unique AP (i.e., *uniqueAPs*) to which these devices were connected. We use the AP connected to as a feature to avoid cases where a user is moving with a phone in power saving mode – in this case, the phone will not generate many network events, but the change of APs show that the user is not sleeping. We assume that a device connects to just a single unique AP when a user is sleeping – this assumption can be relaxed easily (by adding a set of APs into the feature set) in the rare cases where this assumption is not true. In Section 5.1.2, we show that generating these features every one minute struck a good balance between accuracy, computational overhead, and resilience to noise.

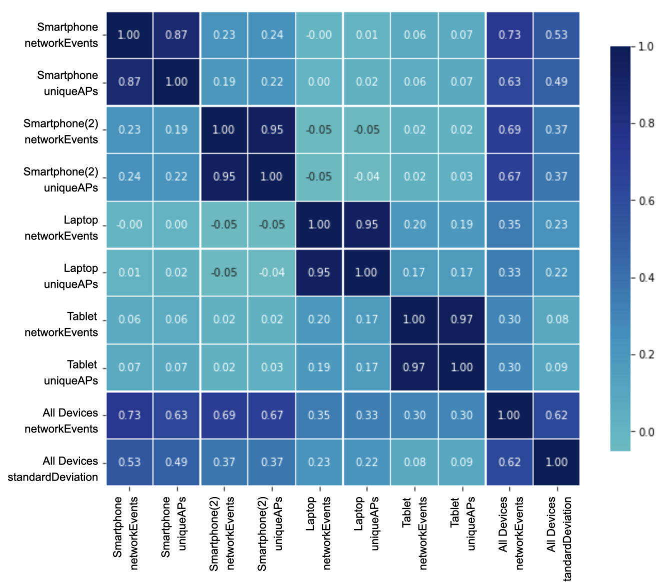

Fig. 4. Correlation heatmap of features extracted from multiple device WiFi network activities.

Figure 4 shows the amount of correlation between sets of features (highly correlated features are darker in color and have scores close to or exactly 1.0). We observe that for all device types (smartphone, laptop, tablet), the number of unique AP associations (uniqueAPs) is highly correlated with the number of WiFi connection events (networkEvents). In Section 5.2, we use these results to remove all highly-correlated features (i.e., unique APs for all devices) and adopt a standard machine learning pipeline of feature selection. *SleepMore*'s final feature selection uses importance weights, which is the number of times a feature is used in the fitted trees inside the Random Forest classifier. As a last step in the pre-processing pipeline, it assigns sleep labels for our algorithm.

We binarize the users' reported nocturnal sleep duration (i.e., Oura ring baseline) into *sleep (1)* or *awake (0)* states at every interval.

3.3 Sleep Prediction Module

Figure 5 provides a high-level visual representation of the inner workings of our technique. It generates a set of attributes from the user's multi-device WiFi device activity logs collected every one minute between Day_1, 6 pm and Day_2, 6 pm. Using the features as input, the system first runs a machine learning algorithm to classify a user's state as *sleep* or *awake*, and calculates the respective prediction confidence. Then, it takes the longest sequences of sleep states to define when the user is sleeping. In doing so, it applies a smoothing function to the sequence of events and calculates the estimation uncertainty based on the number of low-confidence states present in the sequence.

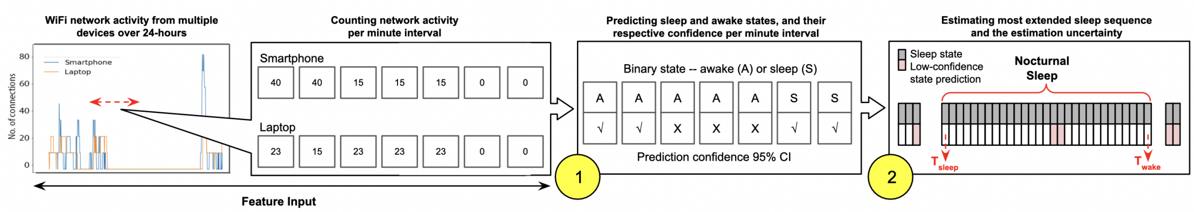

Fig. 5. High-level representation of our technique, receiving WiFi-generated device features as input to an ML classification and estimate sleep through sequence of sleep states.

3.3.1 *Machine Learning Classification with Confidence Interval.*

SleepMore produces a binary predictions of whether the user is currently in a *sleep (1)* or *awake (0)* state. Section 5.1 compares several machine learning techniques, including Naive Bayes (NB), Random Forest (RF), and Extreme Gradient Boosting (XGB) algorithms for this problem. We found that the random forest classifier was the best. Specifically, for *b=1* to *B=200* total trees, it draws a bootstrap sample from the training data. Then, it repetitively grows a tree, T_b, from the bootstrap sample by randomly selecting *d=5* features without replacement and splitting the node using the feature that provides the best split until the desired node size is reached. As the algorithm outputs the total ensemble of all trees to predict a state, $\hat{f}(x)$, it aggregates the prediction by each tree to assign the class label, $\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$, at every one minute interval.

A great advantage to using random forest is that the bagging technique helps reduce the variance of unbiased estimated prediction functions; however, our model must grow deep trees to maintain a low bias and decrease prediction variance. On the other hand, this complexity may lead to less interpretable results, especially in characterizing the statistical distribution of our predictions. To better understand which predictions are our model more confident about, we use the infinitesimal jackknife variance estimation method for bagging proposed by Wager *et al.* [53]. The method takes the covariance with respect to the resampling distribution to generate variance estimates, $\hat{V}_{IJ}$, for all predictions. That is,

$$\hat{V}_{IJ} = \sum_{i=1}^{n} Cov_*[T^*(x), N_i^*]$$

where $T^*(x)$ is the prediction, x, at tree, T, based on the subsample $M_1^*, ..., M_s^*$ of size, s, and N_i^* is the number of times the original sample appears in the resample. This technique would down-weigh each observation by an infinitesimal amount to estimate the variance of a statistic. However, since the variance estimate is calculated with a finite number of trees in practice, it is inherently associated with Monte Carlo error. Our implementation decreases this error by using a large number of trees and subtracting off the Monte Carlo estimate of variance with bias correction. As suggested by Wager *et al.* [53], the unbiased variance estimate, V_{IJ}^{B}, is defined by:

$$V_{IJ}^{B} = \sum_{i=1}^{n} C_i^2 - \frac{s(n-s)}{n} \frac{\hat{v}}{B}$$

where $Ci = \frac{1}{B} \sum_{b=1}^{B} N_{bi}^{*} - s/n)(T_b^{*} - \bar{T}^{*})$, given n is number of training examples, and $\hat{v} = \frac{1}{B} \sum_{b=1}^{B} (T_b^{*} - \bar{T}^{*})^2$. Finally, given the variance estimates, we find the sample mean prediction probability with a 95% confidence interval by randomly sampling 1000 predictions per user. We labeled predictions with low confidence for those outside the upper and lower confidence interval. Random forest predictions are asymptotically normal; however, in setting a sufficiently large number of samples, the theorem holds true even if the distribution of the sample means is not normally distributed.

3.3.2 Sleep Duration Estimation Model. The outcomes from our ML model are a sequence of predicted states, $S = \{s_1, ..., s_m\}$ and their respective confidence, $C = \{c_1, ..., c_m\}$, every 24 hours (m = 1440 minutes). By putting all binary state predictions together, we can conceivably determine a user's nocturnal sleep period for that 24-hour cycle by choosing the most extended occurrence of sleep states (Figure 2A). This nocturnal sleep period could include brief awake states as it is not unusual for a user to wake up once or twice during the night. However, our consideration toward the most extended sleep period, specifically during the night, may perversely overlook short sleep states that can occur between long awake states as in Figure 2B.

Thus, the system component requires solving an estimation problem to accurately identify both a user's bedtime (T_{sleep}) and wake times (T_{wake}) using the series of predicted binary states. We compare two different estimation techniques for this problem: predicting with moving averages (i.e., MVA) and smoothed aggregation (i.e., AGG). Both prediction methods use the hypothesis that a user is more likely to be predicted in a sleep state when their past events assert that they were sleeping and vice versa.

(1) *Predicting with Moving Averages:* When using MVA, *SleepMore* will provide sufficient weight to past states in the predictor to reduce the probability of predicting an awake state when a sudden burst of WiFi activity occurs amid a nocturnal sleep period (e.g., a background app notification). Instead, it will require a series of steadily rising WiFi network device activities to change the prediction from asleep to awake. Similarly, valuing the past awake states as an additional predictor would reduce the model predicting a sleep state when WiFi device activity briefly dips (e.g., the user leaves home for a short time).
(2) *Smoothed Aggregation:* Another approach is by applying smoothed aggregation (AGG) whereby we total the sum of predicted sleep states over a larger observation window. For example, by considering predicted sleep states over a 30 minutes window, we produce a range of prediction states that will likely contain a 95% confidence level of the user's state we are interested in. We smooth out the representation by applying a Savitzky-Golay (SG) filter [43] to determine the most extended sleep period. With this smoothing, even brief sleep episodes, as shown in Figure 2B, will be considered as part of the user's sleep period. Then, we can determine the start of the most extended sleep period as T_{sleep} and the end of the period as T_{wake}.

Our comparison on the performance of these estimation methods found MVA worked best in Section 5.3. We empirically determined the sliding window, $W = 5$, based on the achieving optimal performance. For each sequence of states, we calculate the moving average at time period, t, as:

$$MVA_t = \frac{s_t + s_{t-1} + s_{t-2} + ... + s_{W-(t-1)}}{W}$$

If $s_t, s_{t+1}, s_{t+2}, ..., s_T$ is a sequence of sleep states, our model identifies the longest continuous sequence as the nocturnal sleep period for that 24-hour period. The start of a sleep period corresponds to the first occurrence of sleep states, $T_{sleep} = s_t$. The end of a sleep period corresponds to the last occurrence of a sleep state within the longest sequence $T_{wake} = sT$.

3.3.3 Use of Uncertainty Quantification to Address Noisy Data. Recall in Section 2.4, our technique must overcome system challenges that are mainly attributed to noisy WiFi data; for example, device-specific activities do not reflect the user's actual physical state. To handle spurious events, we implement uncertainty quantification, which carries forward the prediction confidence in our ML model. Specifically, we define all prediction states:

$$c_m = \begin{cases} 0, & \text{if prediction mean within 95\% CI} \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

Accordingly, we calculate the rate of estimation uncertainty instanced by the number of low confidence predictions within the sequence, $C = \{c_1, ..., c_m\}$. As will be discussed in Section 5.3, the degree of tolerable uncertainty can be specified and only days with uncertainty less than that threshold are considered. In our case, our analysis found more than 80% of our sleep estimates within 5% uncertainty rate, thus, regarding sequences with more than 5% uncertainty rate as spurious sequences.

3.3.4 Cloud-based Implementation. Our models are built offline with Python, and the ML component is implemented using the standard *scikit-learn* and *forestci* framework [34, 42]. The ML model is trained with a combination of a general dataset and 40% of the user's data, thus semi-personalized. Its critical parameters, namely the number of trees in the forest and the minimum number of samples required to split an internal node, are tuned to achieve optimal performance and unbiased variance estimates. *SleepMore* is built as a cloud-based web service, where the training of a semi-personalized model and testing of new data are completed in a matter of seconds. However, quantifying the uncertainty of a random forest through the infinitesimal jackknife method is typically more expensive than obtaining the random forest.

4 USER STUDY

To evaluate *SleepMore*, we conducted an IRB-approved study, with all participants providing written informed consent before participating. Table 1 summarizes our participants' data.

4.1 Procedure and Participants

Our primary user study involved 46 undergraduates in on-campus university dormitory housing and lasted for four weeks between March and May 2021. Participants met at the start of the study to collect an Oura ring (gen 2) sleep wearable tracker and a study-issued smartphone for baseline data and attend a one-time in-lab clinical assessment to assess their sleep health. However, the study did not have any inclusion/exclusion criteria, especially their sleep habits. That is to say, participants were not required to report receiving a formal sleep disorder diagnosis (e.g., sleep apnea). They were a mix of habitual and irregular sleepers before they were recruited as participants. Throughout the study, participants were encouraged to utilize their on-campus residence WiFi frequently. We recorded all the MAC addresses of students' personal smartphones (Android or iOS), laptops, and tablets. We use these MAC addresses to extract the RTLS data only for devices owned by study participants. Note: by default, the MAC addresses in the RTLS data are hashed. Thus, we only identify individual users after explicitly providing us with their MAC addresses. At the end of the study, the devices were returned, and each participant received compensation of up to USD 7.40 weekly in cash if they fulfilled all study requirements. These requirements include: (a) regularly wearing the Oura ring and logging their sleep, (b) installing the required data collection apps on their smartphone, and (c) using the campus WiFi while in their dorms.

To further validate the applicability of *SleepMore* in a home setting, we conducted a supplementary study of a private homeowner over four weeks. Unlike student users, the home participant directly provided us with event logs from accessing their home WiFi history in a .csv file and self-reported sleep logs. The logs filtered out other devices but their own.

	Dormitory-living (primary)	**Private Residence** (supplementary)
Users	46 undergraduates	1 homeowner – family dwelling
Age	mean: 22.5, stdev.: 1.7	45
Study Duration	4 weeks per user	4 weeks
Oura baseline	14 - 24 days (mean: 17 days)	-
Sleep summary	Bedtime: 12:00 am - 5:30 am (mean: 1:47 am), Wake time: 5:30 am - 1:15 pm (mean: 8:29 am), Sleep duration: 195 - 660 mins (mean: 401 mins)	Bedtime: 12:00 am Waketime: 6:00 am (weekday), 7:00 am (weekend) Sleep duration: 360 mins
% of time spent sleeping	≈ 12-20% of the day per participant	25% of the day

Table 1. Study data summary

We extracted WiFi network activity data for all user-owned devices for the days when their Oura baseline data and self-reports were correct and available. Note: the distribution of sleep and awake states for each user is highly imbalanced as users spent approximately 20% of the day sleeping.

4.2 Oura Ring Sleep Baseline

Each student participant wore an Oura sleep sensing ring as baseline sleep data to compare against *SleepMore*. The Oura ring estimates sleep using heart rate, HRV (heart rate variability), body temperature, and movement via infrared photoplethysmography (PPG), temperature sensor, and a 3-D accelerometer signals [3]. Participants were instructed to wear the Oura ring (both during the day and night) and sync the data to the Oura app daily. Before the study started, participants were asked to select an Oura ring size that was most comfortable. The Oura app was also installed on their mobile phones, and they were asked to run it daily to sync their sleep data to the cloud. From this cloud data, daily sleep measures such as bedtime, wake time, time-in-bed (TIB), wake after sleep onset (WASO) as well as thirty-second epoch by epoch sleep stages data (wake, light, deep, and REM) were extracted using Oura Health's cloud API. Note: we only use the Oura as a baseline and not as ground truth for medically approved sleep studies requiring highly invasive and expensive polysomnography. Two researchers reviewed the Oura baseline data and identified days with incorrect data to ensure data reliability.

4.3 Ethical Consideration

A key consideration in this work is to preserve users' privacy while still being functional. In practice, WiFi data can reveal many aspects of users' private information. *SleepMore* is designed to keep user privacy in mind by collecting coarse-grain network activity. As such, *SleepMore* only uses network data without knowing the corresponding user activity for accessing WiFi or location information derived from approximating AP locations. In addition, *SleepMore* will only provide user-specific sleep predictions if the user explicitly provides the MAC addresses of their devices. It is important to note that network activity data of student devices were directly collected from the campus infrastructure and bounded by the computing agreements agreed to by each user when they received their WiFi credentials. Without these MAC addresses to user device mappings, *SleepMore* can still produce sleep results at aggregate levels but without attributing results to users. Aggregated results remain helpful as they can provide a good overview of the health status of an entire home or residential dormitory without needing to identify individual users.

5 EVALUATING *SLEEPMORE*

In this section, we evaluate the performance of *SleepMore* using the dataset collected from our primary user study. Through a leave one (user) out cross-validation, we present the accuracy, recall, precision, and F1 scores (weighted average of precision and recall).

5.1 Efficacy of ML Models

Our first evaluation compares the efficacy of different classification algorithms at predicting sleep and awake states in 24 hours. Building on prior findings [26], we contrast the performance of Naive Bayes (NB), Random Forest (RF), and Gradient Boosting (XGBoost), sampling features at every 15 minutes interval. As shown in Table 2, a generalized model using random forest yields the highest accuracy of 79.30% in classifying sleep states.

Algorithm	Accuracy (Acc)	Precision (Prec)	Recall (Rec)	F1	p
Naive Bayes	0.538	0.394	0.833	0.521	p<.01
XGBoost	0.775	0.719	0.425	0.498	p<.01
Random Forest	0.793	0.693	0.582	0.624	-
Random Forest (tuned)	0.798	0.713	0.574	0.625	p<.01

Table 2. Model efficacy with different ML algorithms, and random forest yielding best results.

Train	Acc	Prec	Rec	F1	p
General	0.798	0.713	0.574	0.625	-
+ 10% user	0.825	0.751	0.632	0.681	p<.01
+ 20% user	0.835	0.761	0.666	0.706	p<.01
+ 30% user	0.844	0.771	0.682	0.720	p<.05
+ 40% user	0.851	0.788	0.697	0.736	p<.05
+ 50% user	0.858	0.795	0.714	0.750	p>.1

Table 3. Training data for RF model personalization.

Frequency	Acc	Prec	Rec	F1	p
15 minutes	0.851	0.788	0.697	0.736	-
30 minutes	0.815	0.731	0.625	0.666	p<.01
10 minutes	0.872	0.817	0.746	0.777	p<.05
5 minutes	0.906	0.860	0.824	0.840	p<.01
1 minute	**0.939**	**0.896**	**0.905**	**0.900**	p<.01
45 seconds	0.92	0.876	0.887	0.880	p>.1
30 seconds	0.926	0.875	0.888	0.878	p>.1

Table 4. Sample frequency for RF semi-personalized model.

5.1.1 Model Personalization. Next, we compare the difference between using generalized and personalized models where the personalized models were trained using a portion of the test user's data. Overall, the personalized models improve model performance by a small but significant percentage. As shown in Table 3, using 10% training data to build a per-user model improves model accuracy by ≈ 3% compared to a generalized model (p<.01). Recursively adding 10% of training data slightly improves the model accuracy by ≈ 1%. However, the performance improvements did not achieve a statistical difference beyond 40% training data.

Model personalization also helps to improve recall – the probability that a sleep state is correctly predicted. Figure 6 illustrates the prediction outcomes for user P23 over a night. With 40% training data, recall and overall accuracy improved by 3% and 90% compared to a generalized approach (88%).

5.1.2 Sampling Frequency. Another key input to *SleepMore* is the WiFi network activity data sampling rate. In Section 3.1, we explained that the RTLS data frequency ranges from every 5 seconds to several minutes depending on device use. Guided by prior work [26], we investigate the performance improvements at different WiFi data sampling rates of 30 seconds to 30 minutes. Our analysis, shown in Table 4, found that reducing 15 minutes sampling frequency up to 1 minute significantly improves the overall model performance and recall by ≈ 10%. Further decreasing the sampling rate to 45 or 30 seconds has a slight statistically insignificant performance reduction. Thus, we used a WiFi sampling rate of 1 minute.

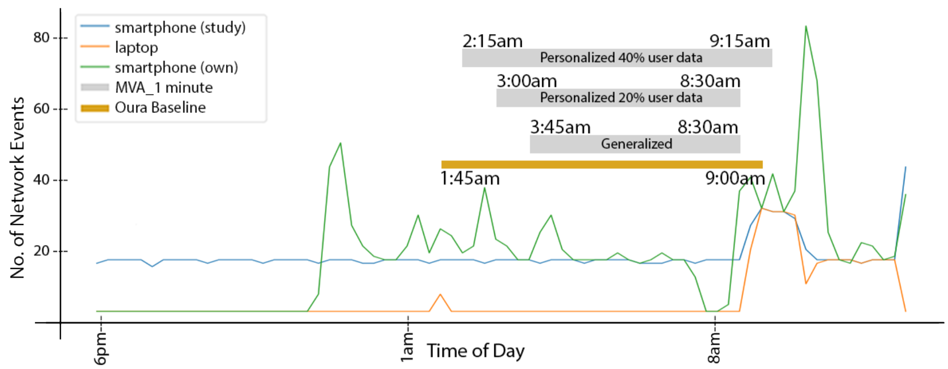

Fig. 6. Personalization improves sleep prediction.

5.1.3 Model Tuning. The inner workings of our chosen classification algorithm random forest use multiple decision trees for decision-making. Fine-tuning our model yields ≈ 94% accuracy, although it is not statistically significant compared to the default model. Several essential parameters for tree-based classification algorithms are the number of trees (n_estimator) and the number of data points placed in a node before it is split (min_samples_split) to improve overfitting. Although minimizing these can avoid overfitting, we iteratively optimized the hyperparameters of our model to reduce bias in calculating the variance estimates *n_estimator=200*, and *min_samples_split=5*.

5.2 Benefit of Using Multiple Devices

A key hypothesis of this work is that using data collected from multiple devices owned by a user will lead to better sleep prediction than using just a single device. To test this, we compare the accuracy of using multiple devices versus just a single preferred primary device. A preferred device is a device (usually a smartphone) that is used the most [52].

Device Type	No. of Users with Device (%)	% WiFi activity per day
Smartphone	46 (100%)	47.30%
Smartphone (study)	13 (28%)	27.10%
Laptop	39 (85%)	35.19%
Tablet	10 (22%)	33.26%

Table 5. Multi-device connection to WiFi per day.

Table 5 summarizes the percentage of time different user-owned devices were connected to the WiFi network. In particular, smartphones were owned by 100% of our users and connected to the monitored WiFi network ≈ 50% of the time. Note: the monitored network was only in the dorms; users were probably elsewhere the rest of the time. Laptops (owned by 85% of our participants) were connected ≈ 35% of the time.

5.2.1 Performance of Multi-Device Features. Figure 7 shows the importance of all features based on our model's fitted trees. The F-score for feature importance is measured in terms of weight, that is, the number of times the feature is used in a tree. Intuitively, a high F-score (1 being the max) reflects how important the feature is independent of other features. This analysis revealed that using the sum of network events across all user-owned devices had the highest F-score compared to using events only from smartphones or tablets.

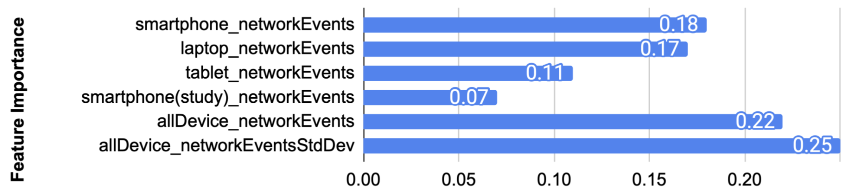

Fig. 7. Feature importance.

Feature	Accuracy	Precision	Recall	F1	p-value
Smartphone	0.893	0.860	0.781	0.812	-
Smartphone + 1 secondary device	0.911	0.851	0.863	0.855	p<.01
Smartphone + multi devices	0.939	0.896	0.905	0.900	p<.01

Table 6. Performance with different device features.

Table 6 compares the model performance with different device features. Using just smartphones yields a 78.10% recall and 89.30% accuracy. However, adding more than two user device features increases recall by more than 10% and overall accuracy by 4% (p<.01). The results strongly suggest that using multiple devices will improve *SleepMore*'s accuracy and recall.

5.3 Efficacy of Sleep Estimation

With the classification component determined, *SleepMore* now has to interpret the classification results. In particular, given a continuous prediction of sleep and awake states, *SleepMore* must now accurately predict the start and end of a user's sleep period (T_{sleep}, T_{wake}) and the sleep duration.

Method	W=5		W=10		W=15		W=20		p-value
	T_{sleep}	T_{wake}	T_{sleep}	T_{wake}	T_{sleep}	T_{wake}	T_{sleep}	T_{wake}	
MVA	39	37	39	37	40	37	41	37	p<.01
AGG	47	42	65	56	80	66	95	78	-

Table 7. Comparison of errors in minutes using two estimation methods on predictions with varying window size (W).

5.3.1 Choice of Smoothing Technique. In Section 3.3.2, we described two techniques to estimate sleep timing: predicting with moving averages (MVA) and smoothed aggregation (AGG). MVA determines a *sleep* or *awake* state by gradually moving the window, defined by the threshold interval, over the predicted outcomes in single increments. Then, we determine the most extended sleep period from the revised outcomes. In AGG, we total the sum of every few sleep states, defined by the threshold interval, and smooth out the representation by applying a Savitzky-Golay (SG) filter [43].

Table 7 tabulates the mean error in users' sleep period with varying window sizes. The errors of T_{sleep} and T_{wake} are calculated as the difference in time between our sleep estimates and the users' Oura ring sleep baseline reference. We observed that using MVA with a rolling window size of 5 to 20 minutes hovered between 39 to 41 minutes in estimating T_{sleep} T_{wake} errors. However, with AGG, errors fluctuate in a more extensive range of 42 to 95 minutes. With *SleepMore* achieving the lowest mean errors in single increments of 5-minutes (regardless of method), we adopt this window size for the MVA estimation method moving forward.

5.3.2 Improving Performance with Uncertainty Rate. These errors thus far include the presence of sleep estimates with significant uncertainty. Recall in Section 3.3.2 that the uncertainty rate is instanced by the number of states predicted with low confidence. To improve the system's performance, we consider only sleep estimates with no more than 5% uncertainty rate, which makes up more than 80% of our prediction samples.

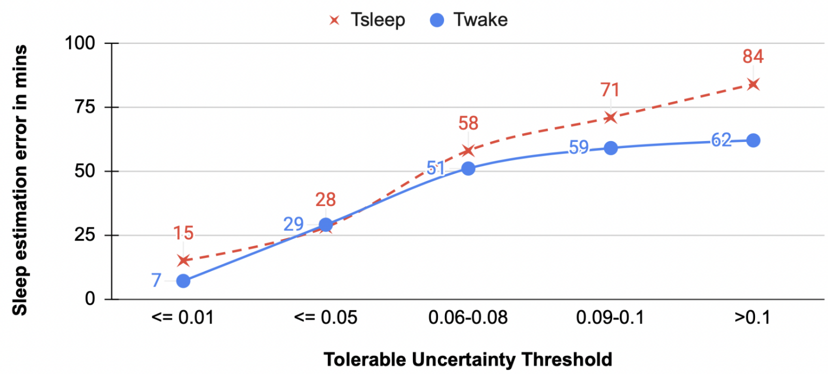

Fig. 8. Sleep estimation error in minutes at varying uncertainty rates.

Figure 8 charts our estimation errors with varying uncertainty rates. Sleep estimations made within 1% uncertainty yielded the lowest errors of 15 minutes for sleep time and 7 minutes for wake times. Estimates made within 5% uncertainty achieved an average error of 28 minutes sleep time and 29 minutes wake time. As the tolerance threshold increases, the technique "filters" out more days with an uncertainty rate higher than the threshold and provides higher accuracy. This mechanism results in a tradeoff where the number of days made by a prediction depends on the tolerance. The higher the tolerance, the more predictions are made at the cost of higher errors. In keeping the threshold within 5% uncertainty, *SleepMore* retains 82% of its prediction outcomes (i.e., 380 out of 460 outcomes of all student participants). Increasing the threshold to 1% only retains 10% of these outcomes.

5.4 Summary of Findings

In its present form, *SleepMore* predicts users' sleep states by employing a fine-tuned semi-personalized (using 40% of a user's data) random forest classification model and infinitesimal jackknife variance estimation method to produce confidence intervals. We collected, on average, 23 days of data per user in our data collection study – thus, nine days' worth of user data would be used for training. Our approach samples WiFi-generated features from multiple user devices every minute. In practice, *SleepMore* does not need to sample WiFi data for the entire day consistently; instead, it just needs to sample during reasonable sleeping periods every day (i.e., at the very least from 8 pm to the following morning on weekdays with afternoons added on weekends). We provide quantitative evidence to strongly suggest that using multiple devices will improve *SleepMore*'s accuracy and recall. Finally, using MVA-5 with a rolling window size of 5 minutes yielded the best results, with sleep errors ranging between 15-28 minutes and wake errors ranging between 7-29 minutes within a 5% uncertainty rate.

6 COMPARATIVE ANALYSIS

This section focuses on the sleep estimates produced by *SleepMore* with different devices and its use in a home setting. We also draw comparisons to understand its core performance with the Oura ring, chosen as a dedicated sleep tracker in our study, under different sleeping patterns and prior prediction techniques.

	Smartphone + multi devices		Smartphone + 1 device		Smartphone only	
	T_{sleep}	T_{Wake}	T_{sleep}	T_{wake}	T_{sleep}	T_{wake}
Median	7	7	18	15	21	27
Mean	17	21	31	32	42	37
Max	182	177	211	403	301	253
Min	0	0	0	0	0	0
Mode	5	0	5	1	5	1
Stdev.	27	32	35	47	49	43
Q1, Q3	5,16	2,26	6,43	4,42	8,58	3,51
UIF, UOF	33, 51	63, 100	99, 156	99, 156	134, 210	123,195

Table 8. Summary statistics of sleep and wake estimation errors with varying devices in minutes.

6.1 Performance with Varying User Devices

Table 8 further breaks down our sleep estimation errors using an MVA-5 method by summarizing the complete statistics with varying devices. We observed that the errors occurring most often were 5 minutes for bedtime and 1 minute for wake time. However, the mean values and standard deviations tend to increase with lesser user devices, reiterating the value of adding more devices to achieve accuracy in sleep monitoring.

A probability density function view of the estimation errors in Figures 9 and 10 also suggest that the outliers were a small fraction of the predictions with more devices used for monitoring. Sleep errors were reduced significantly with more than two secondary devices compared to using one smartphone ($p<.01$) and two devices ($p<.01$). Wake errors did not significantly improve by adding a secondary device compared to using a primary smartphone ($p>.1$) but significantly improved with multiple secondary devices ($p<.01$).

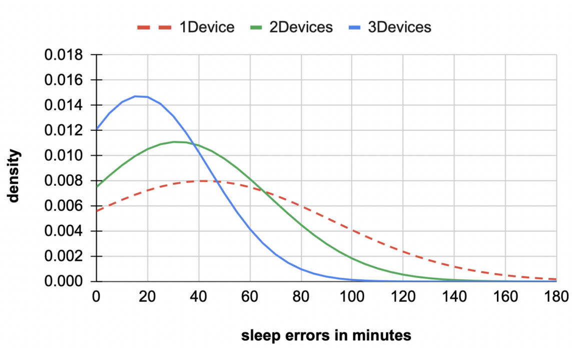

Fig. 9. PDF of sleep error with varying user devices.

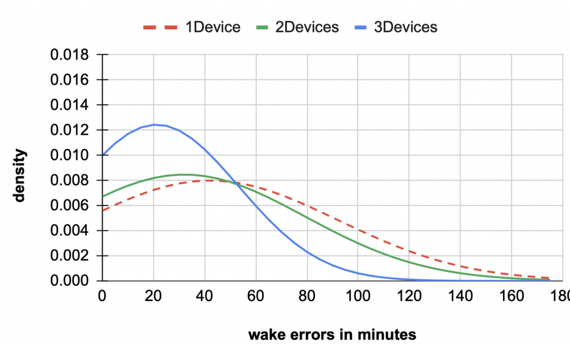

Fig. 10. PDF of wake error with varying user devices.

Characterizing the Outliers: We further analyze the outliers despite monitoring multiple user devices. A primary reason for outliers is the lack of training data, specifically for bedtime hours well past midnight and before dawn. As an example, Figure 11 illustrates *SleepMore* producing significantly inaccurate predictions for two different users, P24 (top) and P27 (bottom), on days they reportedly slept in the early morning hours. Note that these events were each an anomaly of their habitual sleeping habits. It is possible that the lack of training data, especially users sleeping at 4:00 am onwards, had affected the results of these days. We counted only 30 samples (equivalent to less than ≈ 5%) of users' data points who slept beyond 4:00 am throughout the four weeks. Additionally, these data

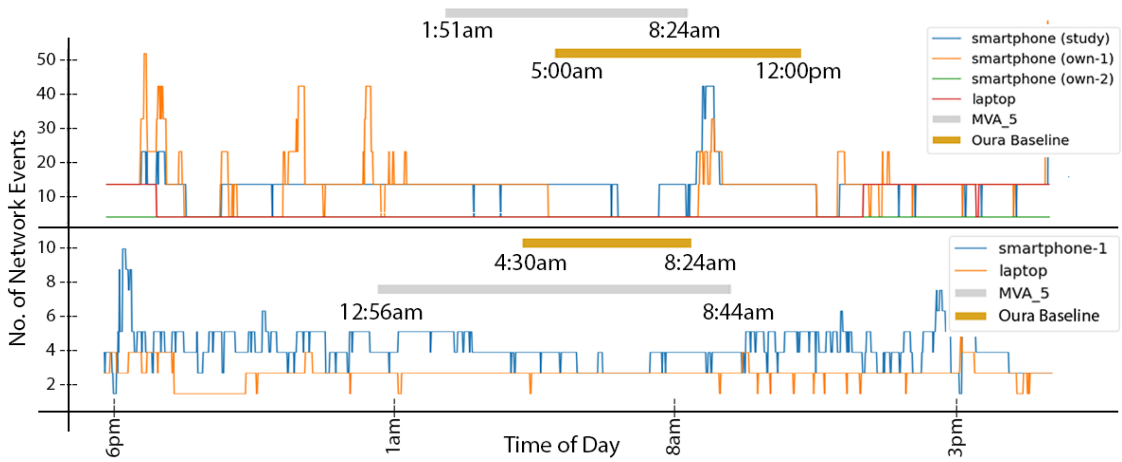

Fig. 11. Larger errors from the lack of training data for bedtime before dawn.

points were mainly attributed to one participant, P37, whose semi-personalized model accurately predicted their irregularities (we discuss this instance in Section 6.2.1). As a result, P24 and P27's semi-personalized models had not learned to predict these outcomes. In similar cases where users did not regularly sleep in the early morning hours, the average bedtime error is within 115 minutes, and the wake time error is within 70 minutes.

Key Takeaway. Indeed, a significant challenge foreseen in our approach is determining a user's bedtime and wake time by monitoring coarse-grained WiFi features. Our estimation method values sleep states made in the last 5 minutes to predict each user's upcoming sleep/awake state and takes the most extended sleep period as the nocturnal sleep duration. This approach assumes that users are present in their homes and connected to their home WiFi for most of their evenings; otherwise, sparse WiFi device activity can negatively impact sleep prediction. Our study currently lacks samples of users sleeping at dawn, thus had affected some outcomes where users were beginning to sleep in the morning. Nevertheless, our approach can successfully reduce these errors with more user devices added for sleep monitoring.

6.2 Contrasting the Oura Ring

Table 9 summarizes the average descriptive statistics for the sleep durations measured by both the Oura ring and predicted by *SleepMore* for the data collection study participants. The differences in these two distributions were statistically insignificant (with p>.1). In particular, both modalities measured that 50% of the participants sleep for at least 400 minutes per day while the total population was measured to sleep for at least 600 minutes, as shown in Figure 12.

6.2.1 Robustness Across User and Sleeping Habits. The following analysis seeks to understand how robust *SleepMore* is to different types of users and their habitual sleep patterns. For each participant, given their set of T_{sleep} and T_{wake} times across multiple days, we can determine the nocturnal sleep duration for that participant.

Figure 13 charts participant P4, who typically slept at 1:25 am (stdev. 36 minutes) every night, corresponding to a *normal* circadian phase. For the entire 12 days that P4 provided useful Oura data, *SleepMore* predicts them getting approximately 7 hours of sleep on average, thus maintaining a regular sleep pattern [41], such as on Day

	$Oura_{duration}$	$SleepMore_{duration}$
Median	404	430
Mean	400	426
Max	680	641
Min	24	210
Mode	428	428
Stdev.	75	67
Q1, Q3	358,448	389,471
UIF, UOF	582,717	594,718

Table 9. Descriptive statistics for Oura ring and *SleepMore*.

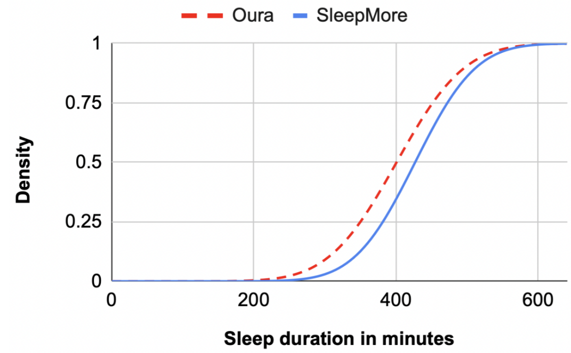

Fig. 12. Sleep duration CDF by Oura and *SleepMore*.

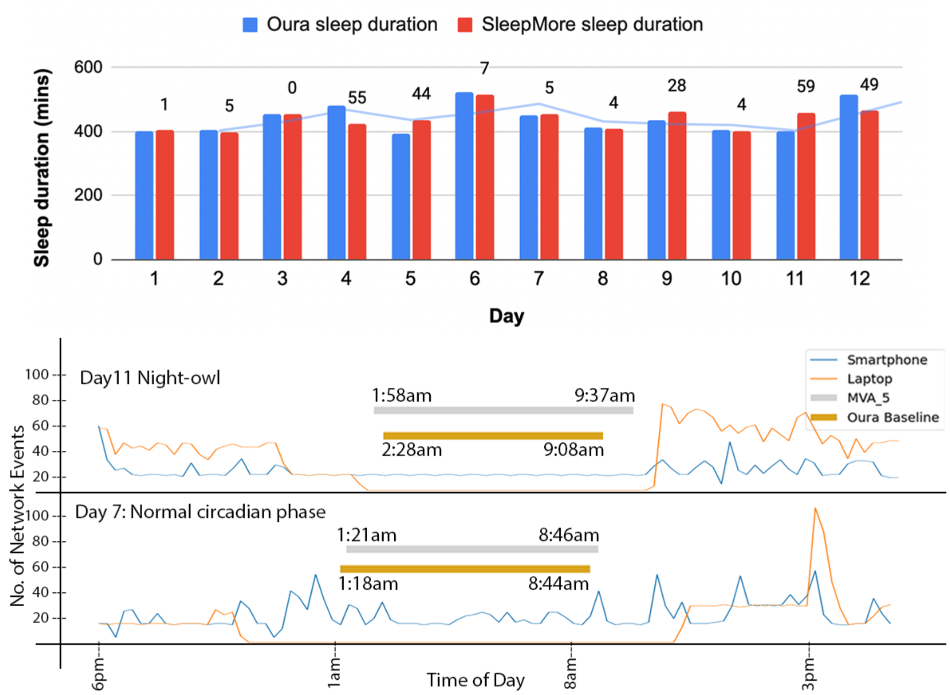

Fig. 13. Comparing Oura and *SleepMore* for P4. Top charts sleep duration difference in minutes.

7. This includes Day 11, when they significantly shifted their usual sleep time – even so, *SleepMore* predicted their sleep duration within 30 minutes sleep time error (and 29 minutes wake time error, totaling 59 minutes).

A separate example is participant P37, as per Figure 14, whose average bedtime clocked at 2:09 am (stdev. 1 hour 08 minutes). On average, participant P37 received 5 hours 30 minutes (stdev. 1 hour 12 minutes) of sleep

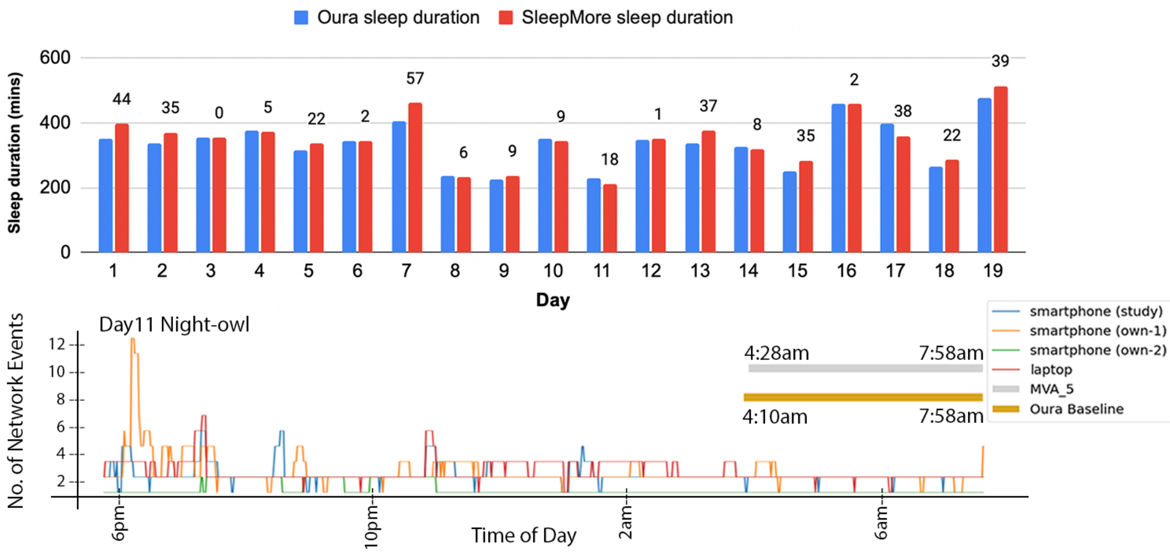

Fig. 14. Night owl sleep schedule by P37, identified to maintain an irregular sleep schedule.

during the study period. Unlike participant P4, we identified participant P37 as an irregular sleeper [47], adopting a mix of habitual *night owl* [28] and *normal* circadian sleep schedules. On Day 11, P37 received approximately 3 hours 30 minutes of sleep. Even with this large shift, *SleepMore* predicted their sleep duration with less than 18 minutes of sleep time error.

Key Takeaway. These examples underscore the robustness of our model in measuring sleep patterns comparable to the Oura ring. The prediction outcomes of bedtime, wake time, and sleep duration are telling of sleep cycles and can provide a fundamental understanding of sleep hygiene at the aggregated and individual level.

6.3 Home-Use Applicability

These evaluations thus far have focused on our primary user study among on-campus student residents. One question that remains unaddressed is *SleepMore*'s performance in a private home setting. Note that sleep baseline data for the homeowner participant is based on their diary logs.

Figure 15 plots the classification of sleep and awake states using their semi-personalized random forest model. Respectively, we plot the standard deviation of each prediction (square root of the variance), giving us a better idea of the disparity of data from one another. As summarized in Table 10, our model achieves an overall performance of 98.6% accuracy and a high 97.5% recall in predicting sleep states. The effect of these predictions leads to our estimation model producing 5 minutes of sleep error and between 8-9 minutes of wake error within a 5% uncertainty rate. Note that all predictions were estimated within a 3% uncertainty rate; therefore, no outliers were removed.

Key Takeaway The primary objective of this second study is to demonstrate the applicability of our system in a private home setting, albeit the sleep baseline data is based on manual entries. It is essential to highlight that our technique is conceptually adaptable to different residential settings as long as users connect their devices to a

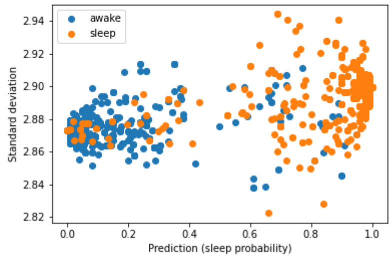

Fig. 15. Classification of sleep states and measure of dispersion for homeowner.

	Homeowner
Acc	0.986
Prec	0.977
Rec	0.975
F1	0.976
T_{sleep}, T_{wake} **(<=1%)**	5, 8 minutes
T_{sleep}, T_{wake} **(<=5%)**	5, 9 minutes

Table 10. *SleepMore*'s performance and errors within 1% and 5% uncertainty rates in a private home setting.

dedicated WiFi network when at home. In this case, we distinguish a user based on their personal devices' MAC addresses and can singularize different occupant's behavior based on their individual list of personal devices.

6.4 Performance Comparison Against Prior Work

Our final analysis compares against prior work that uses similar methods in terms of accuracy. In particular, the prior work also uses WiFi connection data obtained directly from the WiFi infrastructure. The comparison was made against three Bayesian change point detection methods using WiFi device activity from all devices (i.e., *all_networkEvents* feature); normal prior [12], hierarchical prior [10], and an ensemble of normal, uniform, and hierarchical priors [26]. Prior works on sleep detection mentioned in [26] rely on Bayesian change point detection techniques that predict the bedtime and wake-up time, looking at the rate of change of values in a single variable time-series data. Techniques using only a hierarchical prior or a normal prior are useful when users have regular sleep patterns and the data is not subjected to too much noise. However, basic profiling of users is required to best decide on these priors. An ensemble model accommodates the irregular sleepers, including a uniform prior model, the normal prior model, and the hierarchical prior model.

Technique	**Acc**	**Prec**	**Rec**	**F1**	T_{sleep} **error in mins**	T_{wake} **error in mins**
SleepMore	0.939	0.896	0.905	0.900	15-28 (<= 5%)	7-29 (<= 5%)
Ensemble based	0.787	0.586	0.891	0.707	139	175
Norm. Prior	0.799	0.600	0.919	0.726	190	109
Hier. Prior	0.814	0.622	0.914	0.740	193	111

Table 11. Comparison with prior techniques and our sleep estimation errors within 5% uncertainty rate.

Table 11 summarizes the performance comparison based on our primary study (student participants' data) and shows that *SleepMore* achieves significantly better accuracy, recall, and precision (p<.01). A plausible explanation for the change point detection models failing is the absence of location information. Specifically, the amount of WiFi network activity we calculate every minute does not consider changes between places and only checks for a coarse residential location. The above change point detection techniques are suitable in situations where we have no access to the training data and want a high-level understanding of users' behavior in a residential location. However, when we want to do more fine-grained analysis and have access to training data from multiple devices, *SleepMore* can provide more accuracy.

Key Takeaway. While earlier work similarly utilizes WiFi connection data to predict sleep, two main differences set our work apart: First, these works relied solely on single-device monitoring through users' smartphones. Second, change point detection is employed in prior work to determine sleep from changes in WiFi event rates in different locations.

7 DISCUSSION AND LIMITATIONS

Our study's objectives were to develop a sleep prediction solution that can accurately predict sleep by monitoring WiFi device activity for multi-user devices. Here we discuss the implications of our findings.

7.1 Improvements to Handle Corner Cases

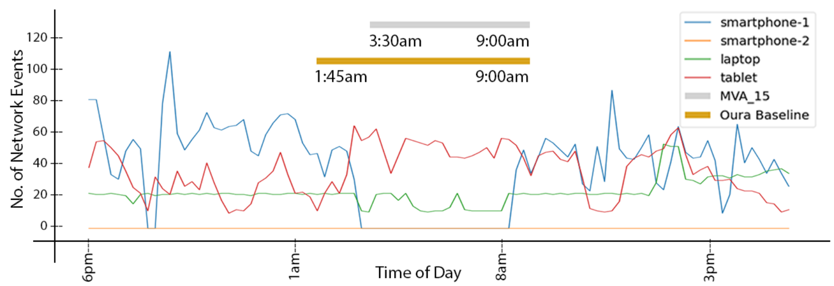

Fig. 16. Outlier with large error.

In situations where Oura predicted a user to sleep before WiFi device activity data is monitored to decrease, our estimation technique will delay predicting the user as asleep. Figure 16 exemplifies an error our system encounters that could be attributed to multiple variables and conditions – it is possible that background activity was running when the user slept (e.g., the device is streaming a video and the user falls asleep). However, it could equally be possible that the Oura was producing erroneous readings. Despite removing sleep predictions with a high uncertainty rate, this is one example that resulted in large errors. Our work continues to explore more sophisticated prediction techniques to handle these corner cases better.

7.2 Assessing Type of Sleep Characteristics

We provide evidence of our technique producing key sleep characteristics comparable to the Oura ring, such as bedtime, wake time, and sleep duration. A clear limitation of utilizing a coarse-grained data source is the inability to differentiate the four sleep stages (i.e., Deep, Light, REM, wake) as obtainable by the Oura ring. Hence, it should be emphasized that our technique aims not to replace existing sleep-sensing modalities that offer fine-grained information but instead to complement the use of such modalities in supporting sleep monitoring, especially over longitudinal periods. At present, the training for our model is notably restricted to nocturnal sleep cycles. With *SleepMore* predicting sleep states every minute, we believe the system can be extended to conceivably detect different sleep patterns, including intermittent wake-ups, sleeping in the morning, or taking daytime naps.

7.3 Shared Information Appliances for Multi-Dwelling Units

Our primary user study of on-campus student residents exemplifies the feasibility of supporting large-scale sleep monitoring in multi-dwelling units. Although the supplementary home study was limited to a single user, it should be noted that our homeowner participant lived in a family dwelling environment. Thus, it is feasible to extend sleep prediction for other family members/occupants by singularizing each user's sleep predictions based on their dedicated list of personal devices. However, our study has limited the monitoring of shared information appliances such as smart TV and smart speakers, which are almost guaranteed to be connected to home WiFi networks. This decision was based on prior research underscoring the dominant utility of smartphones before a user's bedtime and our exploratory analysis discovering the possible utility of other personal devices upon the user waking up. Shared devices may be the first kind that becomes active, thereby more accurately capturing awake states; thus, our work continues to investigate incorporating a wider variety of smart devices to predict sleep more accurately.

8 CONCLUSIONS AND FUTURE WORK

This work proposed *SleepMore* as a promising and practical solution to provide accurate and easy-to-deploy sleep monitoring in residences such as homes and dormitories. *SleepMore* uses a supervised learning and estimation model that leverages WiFi device activity, collected directly from the WiFi infrastructure from multiple user devices to predict sleep. It does this in two steps; First, it determines if a user is in a sleep or awake state every minute using a random forest semi-personalized model. For each state, the model employs an infinitesimal jackknife variance estimation method to calculate the confidence for each prediction, noting down every state with low confidence. Second, it processes these sequences of sleep and awake, using a moving average to estimate the user's bedtime and wake times. It determines the uncertainty rate of a nocturnal sleep estimate instanced by the number of low confidence states present in this sequence. Our validation uses data collected from 46 participants living in on-campus dormitories and one private homeowner. We showed that *SleepMore* produces statistically indistinguishable sleep statistics than the Oura ring (gen 2) and could predict sleep states with a high recall of $\approx 90\%$ recall. This performance renders sleep duration estimations within 15-28 minutes sleep time error and 7-29 minutes wake time error, henceforth proving statistically significant improvements over prior work. A detailed comparative analysis highlighted the importance of multiple user devices to accurately estimate sleep durations and the robustness of *SleepMore* at predicting sleep measures across different users with different sleep patterns and regularities. As future work, we plan to incorporate *SleepMore* into a scalable solution for on-campus health and behavioral risk surveillance to help college students at risk of sleep deprivation and related issues.

REFERENCES

[1] Saeed Abdullah, Mark Matthews, Elizabeth L Murnane, Geri Gay, and Tanzeem Choudhury. 2014. Towards circadian computing: " early to bed and early to rise" makes some of us unhealthy and sleep deprived. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. 673–684.

[2] Bruce M Altevogt, Harvey R Colten, et al. 2006. Sleep disorders and sleep deprivation: an unmet public health problem. (2006).

[3] Marco Altini and Hannu Kinnunen. 2021. The promise of sleep: A multi-sensor approach for accurate sleep stage detection using the oura ring. *Sensors* 21, 13 (2021), 4302.

[4] K Bartel, R Scheeren, and M Gradisar. 2019. Altering adolescents' pre-bedtime phone use to achieve better sleep health. *Health communication* 34, 4 (2019), 456–462.

[5] Francesco P Cappuccio, Daniel Cooper, Lanfranco D'Elia, Pasquale Strazzullo, and Michelle A Miller. 2011. Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies. *European heart journal* 32, 12 (2011), 1484–1492.

[6] Nicholas IYN Chee, Shohreh Ghorbani, Hosein Aghayan Golkashani, Ruth LF Leong, Ju Lynn Ong, and Michael WL Chee. 2021. Multi-night validation of a sleep tracking ring in adolescents compared with a research actigraph and polysomnography. *Nature and Science of Sleep* 13 (2021), 177.

[7] Xianda Chen, Yifei Xiao, Yeming Tang, Julio Fernandez-Mendoza, and Guohong Cao. 2021. ApneaDetector: Detecting Sleep Apnea with Smartwatches. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–22.

[8] Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tanzeem Choudhury, and Andrew T Campbell. 2013. Unobtrusive sleep monitoring using smartphones. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. IEEE, 145–152.

[9] Helen Crompton and Diane Burke. 2018. The use of mobile learning in higher education: A systematic review. *Computers & Education* 123 (2018), 53–64.

[10] Andrea Cuttone, Per Bækgaard, Vedran Sekara, Håkan Jonsson, Jakob Eg Larsen, and Sune Lehmann. 2017. Sensiblesleep: A bayesian model for learning sleep patterns from smartphone events. *PloS one* 12, 1 (2017), e0169901.

[11] Massimiliano De Zambotti, Nicola Cellini, Aimee Goldstone, Ian M Colrain, and Fiona C Baker. 2019. Wearable sleep technology in clinical and research settings. *Medicine and science in sports and exercise* 51, 7 (2019), 1538.

[12] Yassine El-Khadiri, Gabriel Corona, Cédric Rose, and François Charpillet. 2018. Sleep Activity Recognition Using Binary Motion Sensors. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 265–269.

[13] Liese Exelmans and Jan Van den Bulck. 2016. Bedtime mobile phone use and sleep in adults. *Social Science & Medicine* 148 (2016), 93–101.

[14] Julio Fernandez-Mendoza, Fan He, Susan L Calhoun, Alexandros N Vgontzas, Duanping Liao, and Edward O Bixler. 2020. Objective short sleep duration increases the risk of all-cause mortality associated with possible vascular cognitive impairment. *Sleep health* 6, 1 (2020), 71–78.

[15] Weixi Gu, Longfei Shangguan, Zheng Yang, and Yunhao Liu. 2015. Sleep hunter: Towards fine grained sleep stage tracking with smartphones. *IEEE Transactions on Mobile Computing* 15, 6 (2015), 1514–1527.

[16] Lauren Hale, Wendy Troxel, and Daniel J Buysse. 2020. Sleep health: An opportunity for public health to address health equity. *Annual review of public health* 41 (2020), 81–99.

[17] Tian Hao, Guoliang Xing, and Gang Zhou. 2013. isleep: Unobtrusive sleep quality monitoring using smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. 1–14.

[18] Jing-wen He, Zhi-hao Tu, Lei Xiao, Tong Su, and Yun-xiang Tang. 2020. Effect of restricting bedtime mobile phone use on sleep, arousal, mood, and working memory: a randomized pilot trial. *PloS one* 15, 2 (2020), e0228756.

[19] Max Hirshkowitz, Kaitlyn Whiton, Steven M Albert, Cathy Alessi, Oliviero Bruni, Lydia DonCarlos, Nancy Hazen, John Herman, Eliot S Katz, Leila Kheirandish-Gozal, et al. 2015. National Sleep Foundation's sleep time duration recommendations: methodology and results summary. *Sleep health* 1, 1 (2015), 40–43.

[20] Chen-Yu Hsu, Aayush Ahuja, Shichao Yue, Rumen Hristov, Zachary Kabelac, and Dina Katabi. 2017. Zero-effort in-home sleep and insomnia monitoring using radio signals. *Proceedings of the ACM on Interactive, mobile, wearable and ubiquitous technologies* 1, 3 (2017), 1–18.

[21] Zhengjie Huang, Wenjin Wang, and Gerard de Haan. 2021. Nose Breathing or Mouth Breathing? A Thermography-Based New Measurement for Sleep Monitoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3882–3888.

[22] Violaine Kubiszewski, Roger Fontaine, Emmanuel Rusch, and Eric Hazouard. 2014. Association between electronic media use and sleep habits: An eight-day follow-up study. *International Journal of Adolescence and Youth* 19, 3 (2014), 395–407.

[23] Clete Kushida. 2012. *Encyclopedia of sleep.* Academic Press.

[24] Diane S Lauderdale, Kristen L Knutson, Lijing L Yan, Paul J Rathouz, Stephen B Hulley, Steve Sidney, and Kiang Liu. 2006. Objectively measured sleep characteristics among early-middle-aged adults: the CARDIA study. *American journal of epidemiology* 164, 1 (2006), 5–16.

[25] Lorrie Magee and Lauren Hale. 2012. Longitudinal associations between sleep duration and subsequent weight gain: a systematic review. *Sleep medicine reviews* 16, 3 (2012), 231–241.

[26] Priyanka Mary Mammen, Camellia Zakaria, Tergel Molom-Ochir, Amee Trivedi, Prashant Shenoy, and Rajesh Balan. 2021. WiSleep: Scalable Sleep Monitoring and Analytics Using Passive WiFi Sensing. arXiv:2102.03690 [eess.SP]

[27] Stijn AA Massar, Xin Yu Chua, Chun Siong Soon, Alyssa SC Ng, Ju Lynn Ong, Nicholas IYN Chee, Tih Shih Lee, Arko Ghosh, and Michael WL Chee. 2021. Trait-like nocturnal sleep behavior identified by combining wearable, phone-use, and self-report data. *NPJ digital medicine* 4, 1 (2021), 1–10.

[28] Luciano Mecacci and Alberto Zani. 1983. Morningness-eveningness preferences and sleep-waking diary data of morning and evening types in student and worker samples. *Ergonomics* 26, 12 (1983), 1147–1153.

[29] Jun-Ki Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, and Jason I Hong. 2014. Toss'n'turn: smartphone as sleep and sleep quality detector. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 477–486.

[30] Merrill M Mitler, William C Dement, and David F Dinges. 2000. Sleep medicine, public policy, and public health. *Principles and practice of sleep medicine* 2, 4 (2000), 453–462.

[31] A. Networks. [n.d.]. RTLS - integrating with the RTLS data feed. https://community.arubanetworks.com/aruba/attachments/aruba/unified-wired-wireless-access/23715/1/RTLSintegrationv6.docx. Accessed: 2021-11-07.

[32] Maurice M Ohayon, Charles F Reynolds III, and Yves Dauvilliers. 2013. Excessive sleep duration and quality of life. *Annals of neurology* 73, 6 (2013), 785–794.

[33] Oura Health Oy. [n.d.]. New Oura Ring Generation 3. https://ouraring.com

[34] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[35] Geraldine S Perry, Susheel P Patil, and Letitia R Presley-Cantrell. 2013. Raising awareness of sleep as a healthy behavior. *Preventing chronic disease* 10 (2013).

[36] Pew Research Center. April 7, 2021. Internet/Broadband Fact Sheet. https://www.pewresearch.org/internet/fact-sheet/internet-broadband/#who-has-home-broadband

[37] Pew Research Center. April 7, 2021. Mobile Fact Sheet. https://www.pewresearch.org/internet/fact-sheet/mobile/#ownership-of-other-devices

[38] Pew Research Center. December 8, 2020. The promise and pitfalls of using passive data to measure online news consumption. https://www.pewresearch.org/journalism/2020/12/08/the-promise-and-pitfalls-of-using-passive-data-to-measure-online-news-consumption/

[39] Pew Research Center. July 16, 2021. Home broadband adoption, computer ownership vary by race, ethnicity in the U.S. https://www.pewresearch.org/fact-tank/2021/07/16/home-broadband-adoption-computer-ownership-vary-by-race-ethnicity-in-the-u-s/

[40] Pew Research Center. June 3, 2020. Experts Predict More Digital Innovation by 2030 Aimed at Enhancing Democracy. https://www.pewresearch.org/internet/2020/06/30/experts-predict-more-digital-innovation-by-2030-aimed-at-enhancing-democracy/

[41] Andrew JK Phillips, William M Clerx, Conor S O'Brien, Akane Sano, Laura K Barger, Rosalind W Picard, Steven W Lockley, Elizabeth B Klerman, and Charles A Czeisler. 2017. Irregular sleep/wake patterns are associated with poorer academic performance and delayed circadian and sleep/wake timing. *Scientific reports* 7, 1 (2017), 1–13.

[42] Kivan Polimis, Ariel Rokem, and Bryna Hazelton. 2017. Confidence Intervals for Random Forests in Python. *Journal of Open Source Software* 2, 1 (2017).

[43] William H Press and Saul A Teukolsky. 1990. Savitzky-Golay smoothing filters. *Computers in Physics* 4, 6 (1990), 669–672.

[44] Tauhidur Rahman, Alexander T Adams, Ruth Vinisha Ravichandran, Mi Zhang, Shwetak N Patel, Julie A Kientz, and Tanzeem Choudhury. 2015. Dopplesleep: A contactless unobtrusive sleep sensing system using short-range doppler radar. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 39–50.

[45] Mary E Rosenberger, Matthew P Buman, William L Haskell, Michael V McConnell, and Laura L Carstensen. 2016. 24 hours of sleep, sedentary behavior, and physical activity with nine wearable devices. *Medicine and science in sports and exercise* 48, 3 (2016), 457.

[46] Arlene Smaldone, Judy C Honig, and Mary W Byrne. 2007. Sleepless in America: inadequate sleep and relationships to health and well-being of our nation's children. *Pediatrics* 119, Supplement_1 (2007), S29–S37.

[47] John M Taub. 1978. Behavioral and psychophysiological correlates of irregularity in chronic sleep routines. *Biological Psychology* 7, 1-2 (1978), 37–53.

[48] Roxana Tiron, Graeme Lyon, Hannah Kilroy, Ahmed Osman, Nicola Kelly, Niall O'Mahony, Cesar Lopes, Sam Coffey, Stephen McMahon, Michael Wren, et al. 2020. Screening for obstructive sleep apnea with novel hybrid acoustic smartphone app technology. *Journal of Thoracic Disease* 12, 8 (2020), 4476.

[49] Siao Hui Toh, Erin K Howie, Pieter Coenen, and Leon M Straker. 2019. "From the moment I wake up I will use it… every day, very hour": a qualitative study on the patterns of adolescents' mobile touch screen device use from adolescent and parent perspectives. *BMC pediatrics* 19, 1 (2019), 1–16.

[50] Simone Tonti, Brunella Marzolini, Maria Bulgheroni, et al. 2021. Smartphone-Based Passive Sensing for Behavioral and Physical Monitoring in Free-Life Conditions: Technical Usability Study. *JMIR Biomedical Engineering* 6, 2 (2021), e15417.
[51] Amee Trivedi, Jeremy Gummeson, and Prashant Shenoy. 2020. Empirical characterization of mobility of multi-device internet users. *arXiv preprint arXiv:2003.08512* (2020).
[52] Alexander JAM Van Deursen, Colin L Bolle, Sabrina M Hegner, and Piet AM Kommers. 2015. Modeling habitual and addictive smartphone behavior: The role of smartphone usage types, emotional intelligence, social stress, self-regulation, age, and gender. *Computers in human behavior* 45 (2015), 411–420.
[53] Stefan Wager, Trevor Hastie, and Bradley Efron. 2014. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research* 15, 1 (2014), 1625–1651.
[54] Bohan Yu, Yuxiang Wang, Kai Niu, Youwei Zeng, Tao Gu, Leye Wang, Cuntai Guan, and Daqing Zhang. 2021. WiFi-Sleep: Sleep Stage Monitoring Using Commodity Wi-Fi Devices. *IEEE Internet of Things Journal* (2021).
[55] Shichao Yue, Yuzhe Yang, Hao Wang, Hariharan Rahul, and Dina Katabi. 2020. BodyCompass: Monitoring sleep posture with wireless signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–25.
[56] Bing Zhai, Ignacio Perez-Pozuelo, Emma AD Clifton, Joao Palotti, and Yu Guan. 2020. Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–33.