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

The availability of commercial wearable trackers equipped with features to monitor sleep duration and quality has enabled more useful sleep health monitoring applications and analyses. However, much research has reported the challenge of long-term user retention in sleep monitoring through these modalities. Since modern Internet users own multiple mobile devices, our work explores the possibility of employing ubiquitous mobile devices and passive WiFi sensing techniques to predict sleep duration as the fundamental measure to complement long-term sleep monitoring initiatives. In this paper, we propose *SleepMore*, which 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-grained 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. Beyond college campuses, we evaluate *SleepMore* on non-student users of different housing profiles. Our results demonstrate *SleepMore* producing statistically indistinguishable sleep statistics from the Oura ring baseline for predictions made within a 5% uncertainty rate. These errors range between 15-28 minutes for determining sleep time and 7-29 minutes for determining wake time, 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 [21]. However, sleep deficiency is a common public health problem, with approximately 50 to 70 million Americans suffering from chronic sleep disorders [3]. 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 [3, 21]. 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, [8, 12, 48] 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 [12]. To overcome the challenges of continuous sleep monitoring on a longitudinal basis, researchers have explored contactless methods using radio, and radar signals [22, 47]. However, the requirements to instrument building infrastructure limit their scalability. Recently, using WiFi signals such as channel state

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.

(CSI) and backscatter information has been proposed as viable solutions for sleep tracking [59]. This is made possible by both smartphones becoming common across all income levels [42] and WiFi solutions becoming prevalent in public, institutional and residential locations [39, 43]. 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 [9], as well as network-side methods, based on monitoring WiFi network activity of the device [28]. 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 [54]. 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 an accurate but generally accessible sensing approach that monitor users' daily sleeping behavior without changing their routines.* Such functionality primarily contributes as a critical resource to the longitudinal requirement in almost all sleep medicine studies, facilitating complete 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) [10]. 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 [9, 28] 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.

Our biomedical sleep research experts conducted an IRB-approved study over four weeks of an academic semester during the COVID-19 pandemic restriction phase. Accordingly, we evaluate the performance of *SleepMore* among 46 undergraduate students who resided on campus. We also conducted a small-scale user study among three non-student participants to evaluate our system's performance in home settings. As part of the study protocols, student participants wore the *Oura ring* (gen 2) [36] wearable sleep tracker for baseline. They were required to connect their devices to the campus WiFi while in their respective residences. While our study specified no criteria on participants' sleep habits, all users must own multiple personal devices. Participants were a mix of habitual and irregular sleepers. They were asked to provide the MAC addresses of smartphones, laptops, and tablets 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 home participant chose to either provide diary logs or use their personal Fitbit. They provided their WiFi network logs directly to us. 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.
   - *SleepMore* is implemented 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.
   - 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 [28]. Note that predictions within 5% uncertainty rate make up 80% of our predicted outcomes.
   - Our system evaluation is supplemented with results comparing *SleepMore* to prior AI techniques. These comparative analyses conclude conditions under which each technique would thrive in predicting sleep using WiFi network data.

(2) We conduct an extensive experimental evaluation of SleepMore on the student population residing on campus. 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 to accurately estimate users' sleep duration and bed and wake times.
   - We extend our solution deployment to two residential setting. The models for home users are tested in a cross-environment, utilizing students' data as the training set. Our results demonstrate the feasibility of *SleepMore* in real-world settings and with actual residents. Further, we provide insights into using smart home devices that are increasingly present in homes and utilize WiFi connections.
   - 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.

Conclusively, *SleepMore* yields statistically indistinguishable sleep duration predictions compared to the commercialized wearable trackers. However, it is essential to emphasize that our prediction mechanism of estimating sleep duration is a partial function of the entire set of fine-grained features that Oura ring and similar wearable sleep trackers can offer. In situations where sleep studies require regular data logging over an extended period, *SleepMore* is competent as a lightweight supplementary tool without requiring users to wear a sleep tracker or install a dedicated mobile app on their smartphone.

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 [3, 38]. 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 [3, 26]. These studies have noted insufficient sleep among adults as a result of lifestyles and work schedules [3], while adolescents' sleep loss is positively associated with more device use and online activities [49]. 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 [27], quality of life [35], cardiovascular illnesses [7], cognitive impairments [16], to name a few.

The paradigm shift of recognizing sleep as a critical predictor of significant health consequences [18] 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 [3, 38] over longitudinal periods and developing new technologies to accomplish this task [32].

2.2 Sleep Sensing Technologies

Sleep studies in clinical practice generally utilize 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. In a comprehensive study over two months, Massar *et al.* [29] compared and contrasted sleep measurements from a consumer sleep-tracker, smartphone-based ecological momentary assessment, and user-phone interactions of 198 users. Their investigation identified stable interindividual differences in sleep behavior, underscoring the utility of these modalities in characterizing population sleep and peri-sleep behavior. We discuss the pros and cons of these options as follows:

(1) **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," [12] 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 [12, 29, 61].
(2) **Contactless sensing** Low compliance with wearables has motivated the design of contactless sensing approaches. Examples include the use of cameras [23], RF or radar sensing to characterize sleep stages [47] and posture [60]. Despite these advances, RF sensing systems have yet to accelerate commercial adoption, while camera-based solutions are more likely to intrude on privacy interests.
(3) **Smartphone-based sensing** By contrast, smartphone-based sensing had arose as an option from the developing smartphone dependency [40] among everyday users, leading to researchers using the "phone-as-a-sensor". Efforts specific to sleep monitoring include distinguishing respiratory patterns through a microphone [53] or inferring user sleep behavior from monitoring screen interactions and application usage [2, 11, 17, 19, 31]. 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.

These works collectively underscore the advancement in sleep monitoring technologies. While promising, much tension in utilizing these modalities in sleep studies arises from the challenge of user retention. Massar *et al.* describe designing an incentive structure to continuously encourage regular data logging across such modalities over the study period. With many people operating e-devices for minutes to hours before actually intending to sleep, we seek to establish an accurate and generally accessible sensing approach that monitors users' daily sleeping behavior without changing their routines. It is important to emphasize that our work aims not to replace the utilization of existing modalities that can obtain key physiological measures. However, it is positioned as a complementary mechanism to support sleep monitoring scenarios over longitudinal periods.

2.3 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 [28], 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 [40, 41, 55].

(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 [57]. We hypothesize that **including all devices owned by the user will significantly improve the accuracy of sleep detection**.
(2) There is a significant gap between the accuracy of smartphone-based methods and those achieved by sleep wearable trackers, which can achieve 96% recall at detecting sleep [4]. With this result in mind, it is essential that our proposed solution, while using coarse-grained data source, achieves **no statistically significant difference with a wearable sleep tracker** in order to serve its use. In this study, we use the Oura Ring (gen 2) as a representative wearable device for sleep monitoring.

2.4 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 [54] 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.

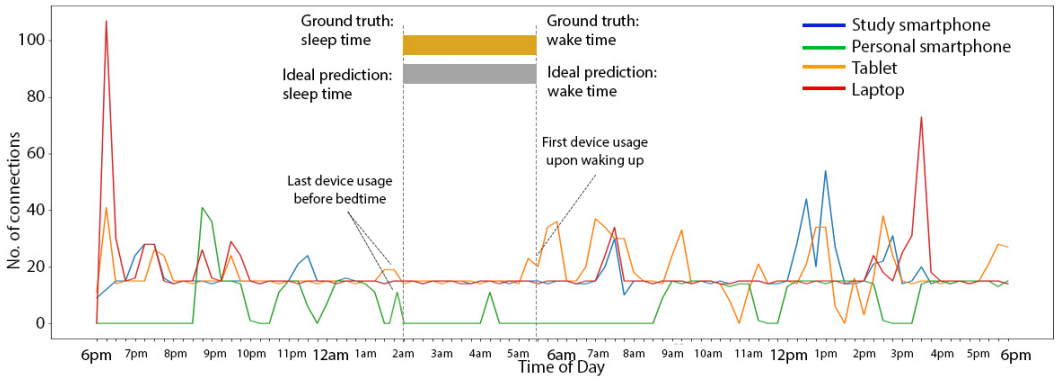

Fig. 1. Utility of smartphone and other devices before sleep and upon waking up. Low WiFi device activity from multiple devices corresponds to a sleep behavior.

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.

2.5 Key Systems Challenges

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

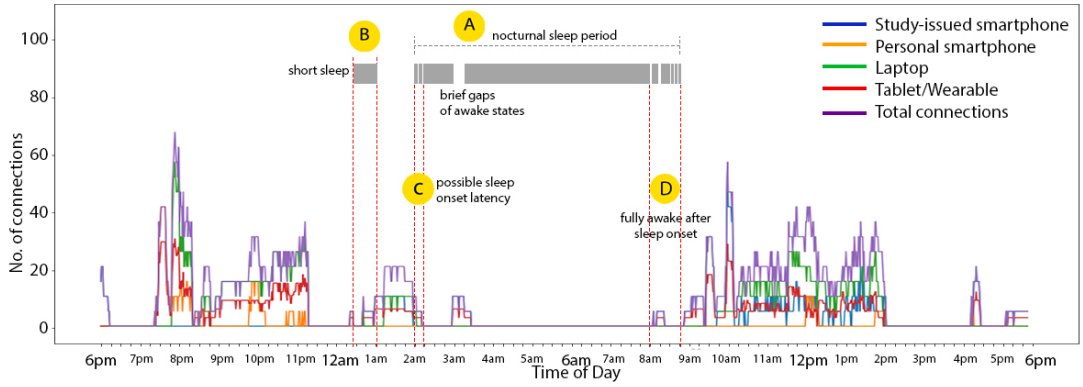

Fig. 2. Challenges predicting sleep with WiFi data.

2.5.1 Noisy Data. Prior work has discussed how noise in WiFi-sensing, such as ping-ponging of AP connections, can negatively impact prediction performance [28]. 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 [15, 24]. 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.

2.5.2 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 [15, 24]. 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 [5, 20, 25]. 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.

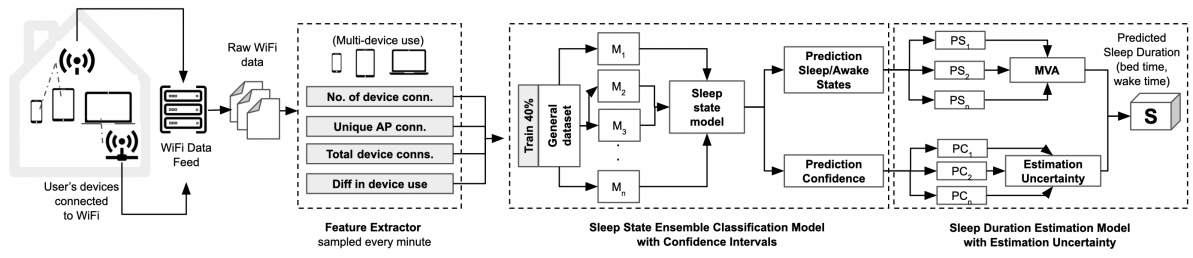

Fig. 3. *SleepMore*, a two-step process in predicting sleep states with confidence interval, and estimating bed and wake times.

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 [33]. 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.

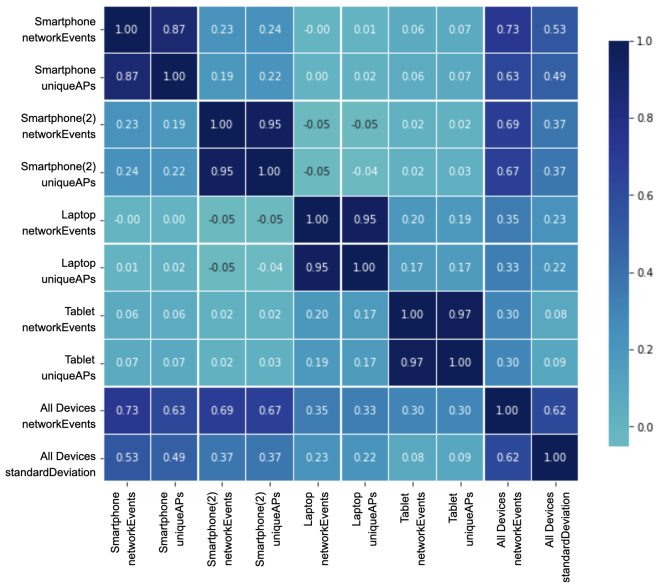

Fig. 4. Correlation heatmap of features extracted from multiple device WiFi network activities.

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.

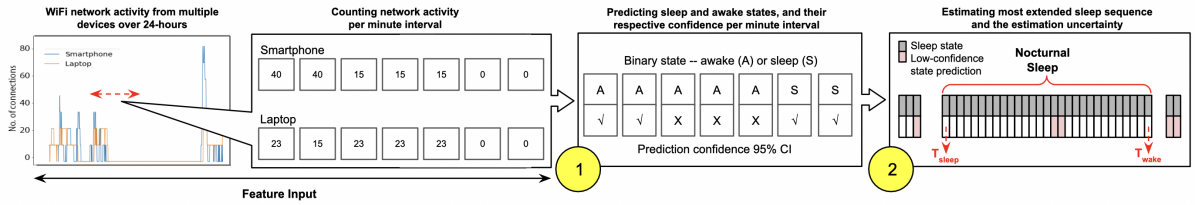

Fig. 5. High-level representation of our technique, receiving WiFi-generated device features as input to an ML classification and estimate sleep through sequence of sleep states.

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), Extreme Gradient Boosting (XGB), and Long Short Term Memory (LSTM) algorithms for this problem. We found that the RF and LSTM yield best results.

In RF, 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. This complexity may lead to less interpretable results, especially in characterizing the statistical distribution of our predictions.

On the other hand, a 2-layer LSTM demonstrated comparable results to RF. The LSTM incorporates three different matrices inside the recurrence, acting as a gating mechanism that allows for information flow from the previous timestep as a function of the current one; an input gate (i_t), a forget gate (f_t) and an output gate (o_t).

Here, w and b represent the weight and biases for gate, x, neurons. h_{t-1} represents the output of the previous LSTM block.

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$
$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$
$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$

These representations in a cell state, c_t are combined with the memory vector for the current timestamp, $\tilde{c_t}$, to produce the hidden representation, h_t. Our final layer is a fully connected layer with a *sigmoid* activation and one neuron. Where W is the weight connection matrix of the layer, the dense layer outputs, y_t, the probability of the event to be of state *sleep* or *awake*.

$$\tilde{c_t} = tanh(w_c[h_{t-1}, x_t] + b_c)$$
$$c_t = f_t * c_{t-1} + i_t * \tilde{c_t}$$
$$h_t = o_t * tanh(c^t)$$
$$y_t = (h_t * W) + b$$

Infinitesimal Jackknife Variance Estimation. To better understand which predictions are our model more confident about, we use the infinitesimal jackknife variance estimation method proposed by Wager *et al.* [58]. 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.* [58], 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 $C_i = \frac{1}{B} \sum_{b=1}^{B} (N_{bi}^* - s/n)(T_b^* - \bar{T}^*)$, 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.

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 [46] 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:

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

If $s_t, s_{t+1}, s_{t+2}, \ldots, s_T$ 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 2.5, 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, $C = \{c_1, \ldots, c_m\}$. 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 approximately 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 with Python, and the ML component is implemented using Keras with TensorFlow backend, *scikit-learn* and *forestci* libraries [1, 37, 45]. The ML model is trained with a combination of a general dataset and 40% of the user's data, thus semi-personalized. *SleepMore* is built as a cloud-based web service, where the training of a semi-personalized model and testing of new data using random forest are completed in a matter of seconds. However, quantifying the uncertainty through the infinitesimal jackknife method is typically more expensive than obtaining the random forest.

As our results will show in Section 5, the performance improvement using LSTM did not reach a significant difference compared to random forest. While LSTM achieves a 3% increase, this improvement is at the expense of

precision and overall accuracy. Since random forest is less computationally expensive, our solution is best suited to random forest with infinitesimal jackknife method. Nonetheless, it should be noted that LSTM as another classification technique is applicable.

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. Participants were recruited in batches between March-May 2021, and data collection was 4-weeks for each batch. The study took place during the COVID-19 pandemic. During this time, there were some restrictions regarding group size gatherings. As part of these restrictions, students attended their lessons online while living in their on-campus dormitories.

In designing this study, our expert sleep researchers fully considered altering the procedure to its bare minimum. It should be noted that sleep studies with gold standard polysomnography (PSG) are typically conducted under conditions where the user's sleep is not interfered with. Having a diagnosed sleep disorder was an exclusion criterion during recruitment; however, there were no criteria for habitual sleep duration or times. Participants included both regular and irregular sleepers with varying sleep habits. Participants were required to own multiple devices, install a dedicated data logging app on their smartphone and actively maintain the diary log. With preserving user privacy in mind, log entries are limited to collecting only sleep and wake times, not reporting wakeful activities at night.

Users, N	46 (23 Male, 23 Female)	**Batch, n**	#1 = 6, #2 = 8, #3 = 7, #4: 6, #5 = 6, #6 = 13
Age	20 - 28 years old, mean: 21.95, stdev.: 1.43	**Year**	44 undergraduates, 2 graduates
Study duration	4 weeks per user	**% of time spent asleep**	≈ 12-20% of the day per student
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)	**Residence**	Self-contained student housing estate, with 30 blocks segregated into 7 residences
Oura baseline	14 - 24 days (mean: 17 days)	**Devices tracked**	WiFi network activity data of 1. smartphones, 2. laptop, 3. tablet

Table 1. Data summary of student participants in the primary user study.

Upon consent, participants met at the start of the study to collect an Oura ring (gen 2) sleep wearable tracker for baseline sleep data; their participation was staggered in six batches. Participants also attended a one-time in-lab clinical assessment to assess their sleep health. Following standard sleep study procedures, participants fulfill a list of questionnaires, including the Beck's Anxiety Inventory [50], Beck's Depression Inventory [51], and Berlin Questionnaire [34]. These questions were as sanity checks for us to be aware of their health states. All participants were in good standing in their mental health self-reports. Similarly, we identified no students with sleep apnea based on the Berlin questionnaire. 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. For students who owned iOS-based smartphones, they received a study-issued Android-based smartphone with a preinstalled data logging app. We use the WiFi MAC addresses of their devices to extract the RTLS data of our 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, all study-issued devices were returned, and each participant received compensation of up to USD 18.12 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 logging app on their smartphone, and (c) using the campus WiFi while in their dorms.

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.

	$HomeUser_1$	$HomeUser_2$	$HomeUser_3$
User	Male, 45	Female, 35	Male, 46
Residence	Single-family house	Apartment	
Household	Family with children	Couple no children	
Study duration	1 month	1 week	
Devices tracked	smartphone, laptop	smartphone (personal), laptop (personal), smartTV, smart-speaker	
Sleep baseline	Diary logs	Fitbit + Diary logs + lights off log	
Sleep summary	Bedtime: 12:00 am Waketime: 6:00 am (weekday), 7:00 am (weekend) Sleep duration: 360 mins	Bedtime: 11:20 pm -12:45 am (mean: 11:46 pm) Waketime: 5:30 - 8:00 am (mean: 6:39 am) Sleep duration: 290 - 480 mins (mean: 376 mins)	Bedtime: 11:15 pm - 12:30 am (mean: 11:37 pm) Waketime: 5:40 - 8:00 am (mean: 6:45 am) Sleep duration: 310 - 450 mins (mean: 372 mins)
Lights off	-	Weekday: 11:00pm, Weekend: 11:30 pm	
% of time spent sleeping	25% of the day	26% of the day	25% of the day

Table 2. Data summary of home participants in the supplementary user study.

To demonstrate the applicability of our approach among other populations, our study expands to include non-student users living in different private home settings, albeit on a smaller scale. The sleep baseline data for this set of participants is based on their preferred choice of a diary log and/or (their personal) Fitbit device. These logging modalities have also proven reliable for sleep monitoring purposes [12]. Fitbit provides similar functionalities and data files as the Oura ring. Unlike student users, home participants directly provided us with their WiFi event logs in a .csv file. Additionally, $HomeUser_2$ provided the logs of their share smart home devices. We summarize the demographic profiles of our home participants in Table 2.

4.2 Sleep Baseline

Each student participant wore an Oura sleep sensing ring as baseline sleep data to compare against *SleepMore*. Students were asked to select an Oura ring size that was most comfortable. Our home participants used their personal Fitbit, also equipped with similar features. These wearables estimate sleep using heart rate, HRV (heart rate variability), body temperature, and movement via infrared photoplethysmography (PPG), temperature sensor, and a 3-D accelerometer signals [4]. Participants were instructed to wear their devices (both during the day and night) and sync the data to the Oura/Fitbit app daily. 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 the respective cloud API. Note: we only use the wearables as a baseline and not as ground truth for medically approved sleep studies requiring highly invasive and expensive polysomnography.

To ensure the baseline sleep data was correct and reliable, we used the sleep summary file containing only duration and time information and the hypnogram file containing wake and sleep stages every 30 seconds. Additionally, we used self-reports to check the alignment of sleep and activity data with self-reports.

Practical Challenges. First, insufficient battery power and missed opportunities to wear the device often led to inaccurate sleep estimation among our students – this is a common challenge faced in many sleep studies [3, 13, 29]. Second, the Oura ring was not equipped with a nap detection feature at the time of the study. A common but small observation among our students was that some morning naps were regarded as a continuation of nocturnal sleep (mismatch between sleep summary and hypnogram). Due to these discrepancies and inadequate ground truth to validate these occurrences, we made the operational decision to exclude training our model for nap-like sleep. Approximately 10% of all data points exhibited this pattern. Hence, we excluded data if there is a mismatch in sleep timings between these three resources. Given a month's participation per student, the data exclusion rate from inaccurate reportings ranged between 7% to the highest of 50%. We excluded no data from home users.

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. We conduct our model evaluation through a leave-one(user)-out cross-validation. As will be explained later, we build a semi-personalized model where a model for each user is trained using 40% of their data. Using all but the user's data as part of the training set, we repeat this process five times, where each training time includes a different set of random samples (of the test users) to build the semi-personalized model. 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 algorithms at predicting sleep and awake states in 24 hours. Building on prior findings [28], 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 3, a generalized model using random forest yields the highest accuracy of 79.30% in classifying sleep states.

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 4, 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.

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

Table 3. 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 4. 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 5. Sample frequency for RF semi-personalized model.

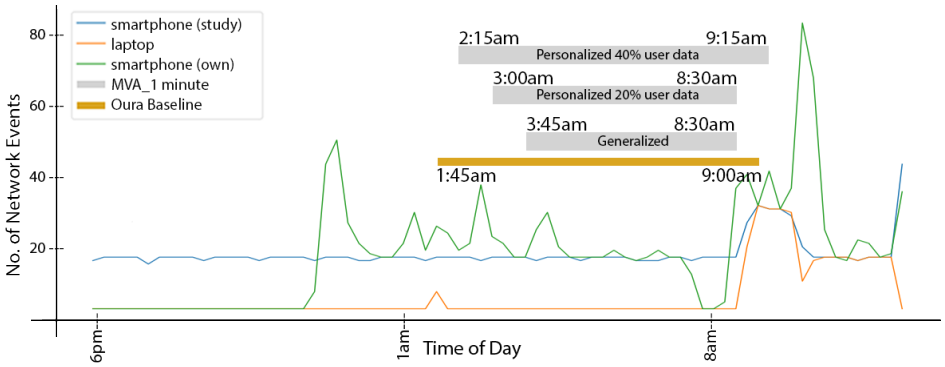

Fig. 6. Personalization improves sleep prediction.

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 [28], we investigate the performance improvements at different WiFi data sampling rates of 30 seconds to 30 minutes. Our analysis, shown in Table 5, 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.

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

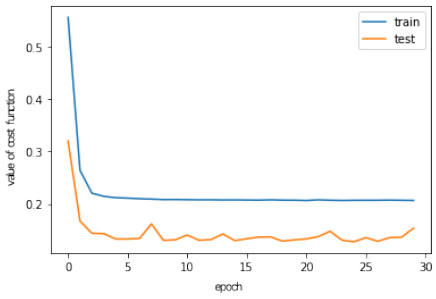

Fig. 7. Value of cost function decreases with epoch size for P10.

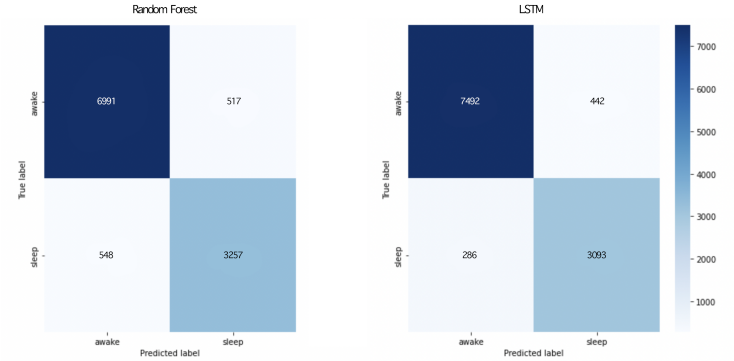

Fig. 8. Confusion matrix comparison between RF and LSTM for P10.

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 using grid search to reduce bias in calculating the variance estimates *n_estimator=200*, and *min_samples_split=5*.

5.1.4 Efficacy of DL model. Our implementation of LSTM begins with a single layer and progressively adding layers to achieve a robust model. Similarly, we used the ten features described in Section 3.2, sampled at 1-minute interval, as the input to the model, and incrementally increased the adjacent samples (i.e., timestep) from 5 to 30 minutes intervals to predict each instance. For example, we set the "lookback period" to every 10 minutes of data in the past to predict the upcoming sample.

Model	Accuracy (Acc)	Precision (Prec)	Recall (Rec)	F1	p
LSTM-5T	0.859	0.841	0.602	0.702	-
LSTM-10T	0.832	0.763	0.676	0.717	p<.01
LSTM-15T	0.858	0.791	0.661	0.720	p<.01
LSTM-30T	**0.928**	**0.879**	**0.883**	**0.881**	p<.01
LSTM-2	**0.930**	**0.863**	**0.913**	**0.887**	p>.1
LSTM-3	0.921	0.866	0.876	0.872	p>.1

Table 6. Comparing *SleepMore* with LSTM network.

Table 6 summarizes the model performance, first, on a single-layer LSTM model with ten hidden layers, 32 samples in each mini-batch, and an epoch size of 5, with varying timesteps. We set the dropout and recurrent dropout rates at 20% as a starting point. Where recall is prioritized, the model yields the best results of 88.3% recall and 92.8% accuracy at 30 minutes time step.

The subsequent addition of an LSTM layer (i.e., LSTM-2) and fine-tuning improve recall. However, we quickly observed declining performance with a third LSTM layer. As the LSTM model is slower to train, we implement a grid search for LSTM-2, attending to the following hyperparameters: *lstm* and *recurrent* dropout rates, *epoch*, and *batch size*. With the optimal hyperparameters at *lstm_dropout=0.01*, *recurrent_dropout=0.2*, *batch_size=64*, and *epoch=5*, a two-layer LSTM model yields best performance at 91.3% recall, increasing recall by 3% compared to a single-layer LSTM. However, this difference is not significant. Figure 7 charts the value of cost function, that is, the error between predicted values and true values, over the entire training set for P10. Contrasting model performance with RF, we plot the confusion matrix for the same user in Figure 8.

While promising, the average performance improvement using LSTM was only on the recall measure and is not significantly different from the RF algorithm. This minimal improvement did not justify building a complex model that is also computationally expensive compared to an RF model. Many applications utilizing LSTM networks have proven successful on large amounts of training data (e.g., months and years) to detect patterns of residential and student life behaviors [6, 56]. Until our application has collected much more data to build classifiers that compare the upcoming sleep duration for the day to the sleep histories of the same days in previous weeks/months, we reckon using the RF algorithm at this stage for our system to function adequately. We discuss our continuing effort for sleep prediction using LSTM in Section 7. Note that our results are based on the RF algorithm in the experiments moving forward.

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 [57].

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 7. Multi-device connection to WiFi per day.

Table 7 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 9 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.

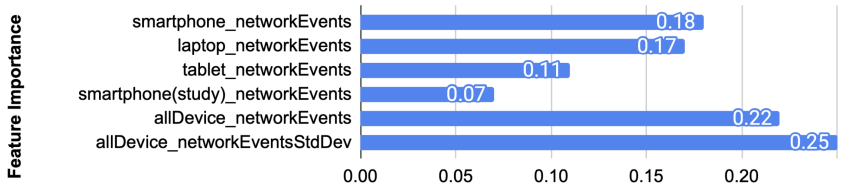

Fig. 9. Feature importance.

Table 8 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.

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 8. Performance with different device features.

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.

5.3.1 Choice of Smoothing Technique. Unfortunately, coupling the confidence interval for each predicted outcome does not provide a straightforward solution to determining a user's sleep duration. For example, a typical representation is depicted in Figure 10 where we observed most sleep states with low confidence spuriously occurring between 1:00 am, and 7:00 am. Ignoring the first and last occurrence of low confidence predictions does not directly translate to estimating the start and end of a sleep period. For these reasons, we described two techniques to estimate sleep timing based on the most extended sleep period in Section 3.3.2: predicting with moving averages (MVA) and smoothed aggregation (AGG). Instead, confidence values will be used to quantify the uncertainty rate of an entire sleep duration prediction once it is estimated.

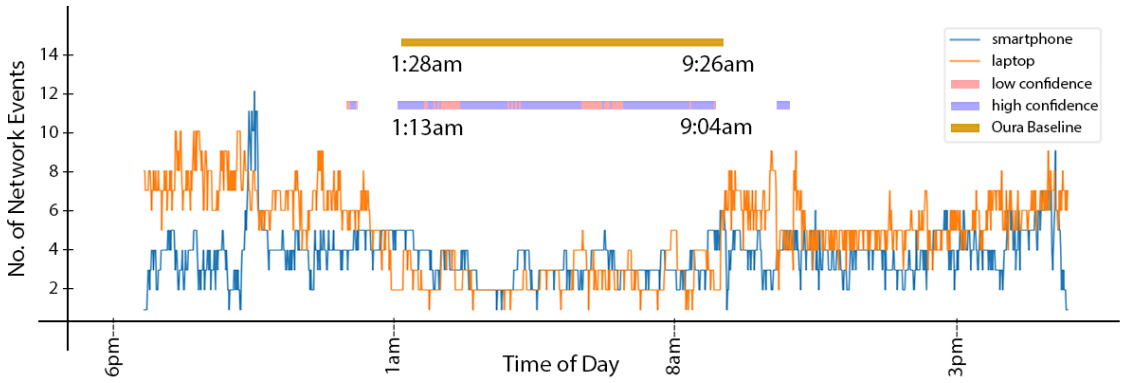

Fig. 10. Predicted states with confidence level for P10.

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 [46].

Table 9 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.

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

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 roughly 80% of our prediction samples.

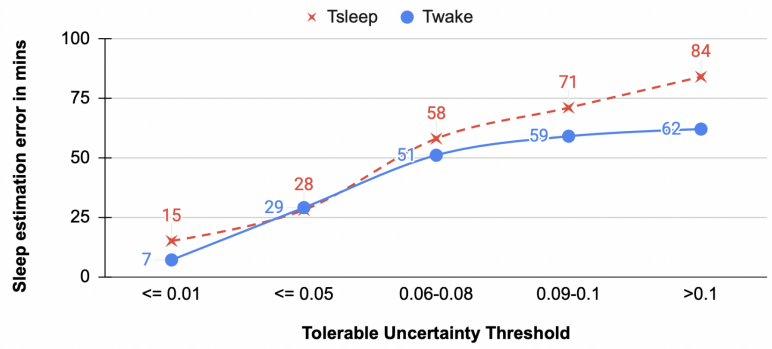

Fig. 11. Sleep estimation error in minutes at varying uncertainty rates.

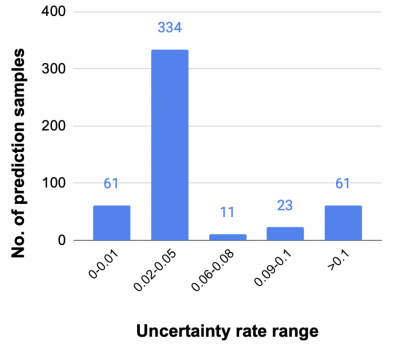

Fig. 12. Prediction sample distribution.

Figure 11 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 the cost of higher errors.

In keeping the threshold within 5% uncertainty, *SleepMore* retains 80.6% of its prediction outcomes. As per Figure 12, most of our predictions are made within 0.02-0.05 uncertainty rate range, thus, excluding 95 of 490 outcomes. Restricting the threshold to 1% retains 12% 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, 17 days of data per user in our data collection study – thus, six 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 in estimating the longest sleep

period. The sleep errors range between 15-28 minutes and wake errors range 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, including the Oura ring, chosen as a dedicated sleep tracker in our primary study. We draw comparisons to understand its performance with varying sleep behaviors and tested our system in a home setting and prior work utilizing AI techniques.

6.1 *SleepMore* and Smart Devices

6.1.1 WiFi Data from Multiple Personal Devices. Table 10 breaks down our sleep estimation errors using an MVA-5 method by summarizing the complete statistics with WiFi data from varying personal 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.

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

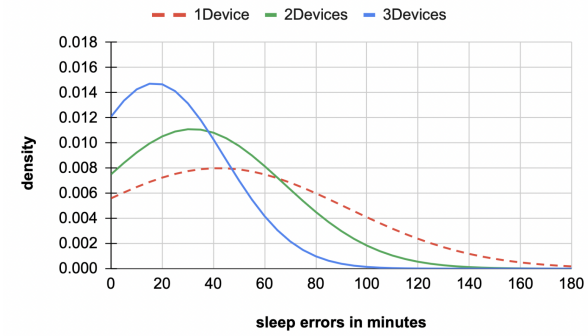

Fig. 13. PDF of sleep error with varying user devices.

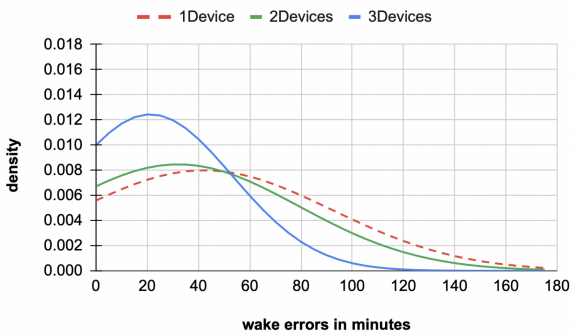

Fig. 14. PDF of wake error with varying user devices.

A probability density function view of the estimation errors in Figures 13 and 14 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 ($p<.01$).

6.1.2 Contrasting the Oura Ring. Table 11 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 15.

Key Takeaway. For *SleepMore* to remain useful as a complementary tool in long-term sleep monitoring, our model performance must match that of the Oura ring. Our findings recommend using multiple personal devices for the system to produce the most accurate predictions.

	$Oura_{duration}$	$SleepMore_{duration}$
Median	404	430
Mean	400	426
Max	680	641
Min	240	210
Mode	428	428
Stdev.	75	67
Q1, Q3	358,448	389,471
UIF, UOF	582,717	594,718

Table 11. Descriptive statistics for Oura ring and *SleepMore*.

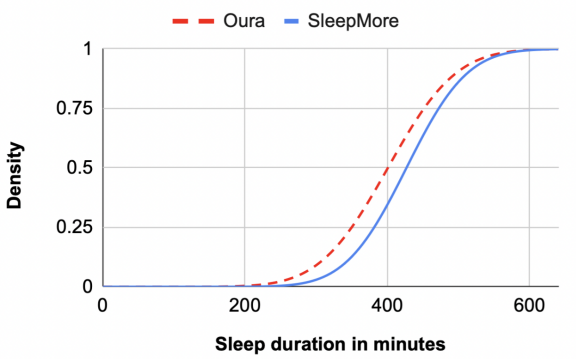

Fig. 15. Sleep duration CDF by Oura and *SleepMore*.

6.2 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. As stated in Table 1, our student participants' average bed and wake times are 1:47 am and 8:29 am, respectively. However, there were outlying records of sleep behaviors where students reportedly went to bed at 5:30 am and woke up at 1:15 pm.

6.2.1 Circadian and Night Owl Phase. Figure 17 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 [44], such as on Day 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 16, 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 during the study period. Unlike participant P4, we identified participant P37 as an irregular sleeper [52], adopting a mix of habitual *night owl* [30] 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.

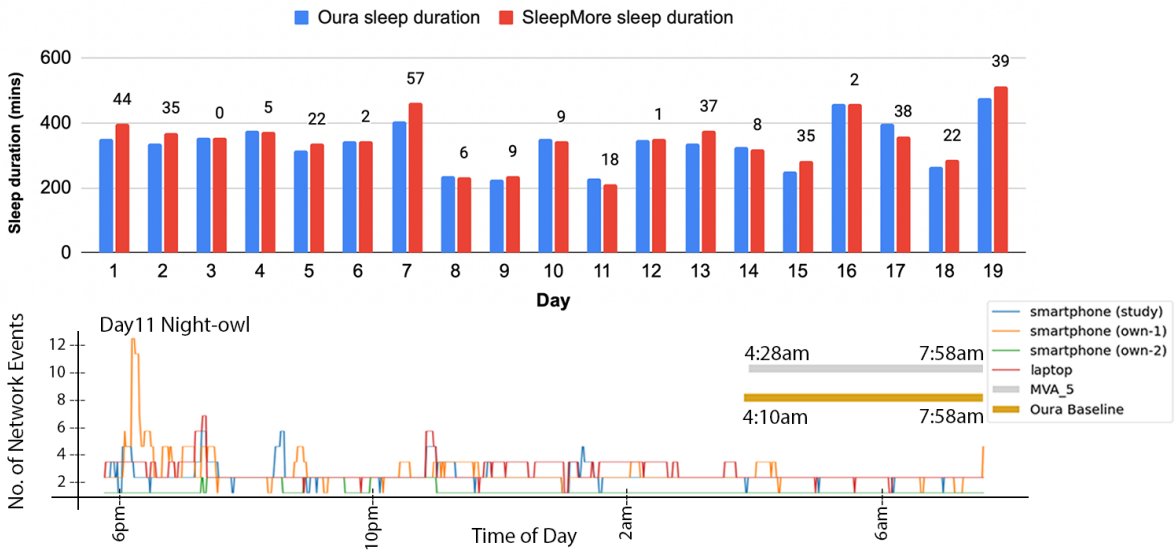

Fig. 16. Night owl sleep schedule by P37, identified to maintain an irregular sleep schedule.

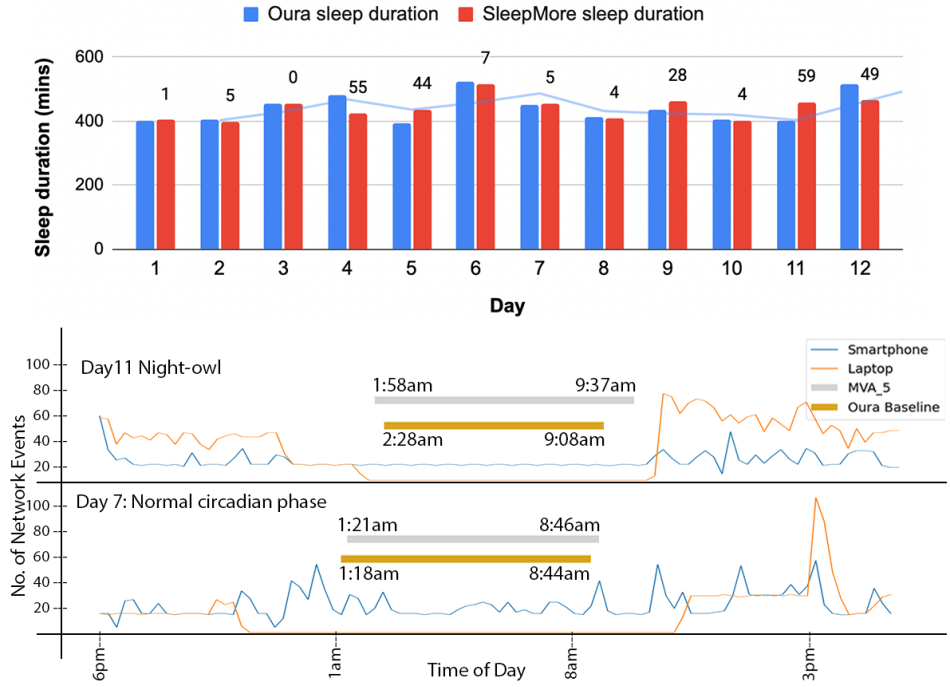

Fig. 17. Comparing Oura and *SleepMore* for P4. Top charts sleep duration difference in minutes.

6.2.2 Disruptive Sleep Event in an Irregular Sleeping Pattern. Figures 18 and 19 show the difference in our model performance for two users, P24 and P27. Note that the disruptive sleep events (bottom chart) were each an anomaly of their habitual sleeping habits. *SleepMore* can accurately predict sleep on days that users display regular sleep patterns. However, a disruptive event that leads to the user sleeping late would cause significant errors at this stage.

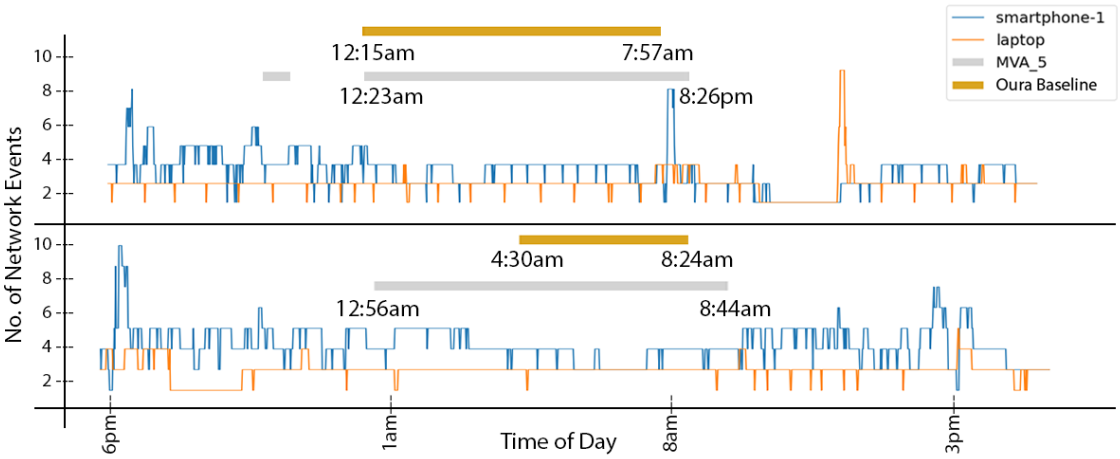

Fig. 18. Model performance is affected by disruptive sleep event (bottom chart) for a regular sleeper, P24.

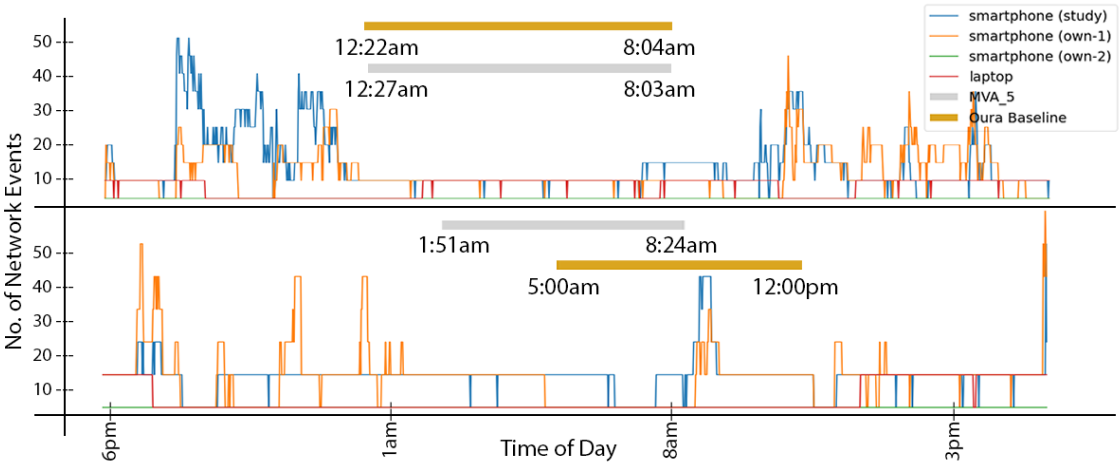

Fig. 19. Model performance is affected by disruptive sleep event (bottom chart) for a regular sleeper, P27.

A primary reason for outliers is the lack of training data, specifically for bedtime hours well past midnight and before dawn. We counted only 30 samples (equivalent to less than ≈ 5%) of (all) students' data points who slept beyond 4:00 am throughout the four weeks. Additionally, these data points were mainly attributed to P37 (above), whose semi-personalized model accurately predicted their irregularities. As a result, P24 and P27's

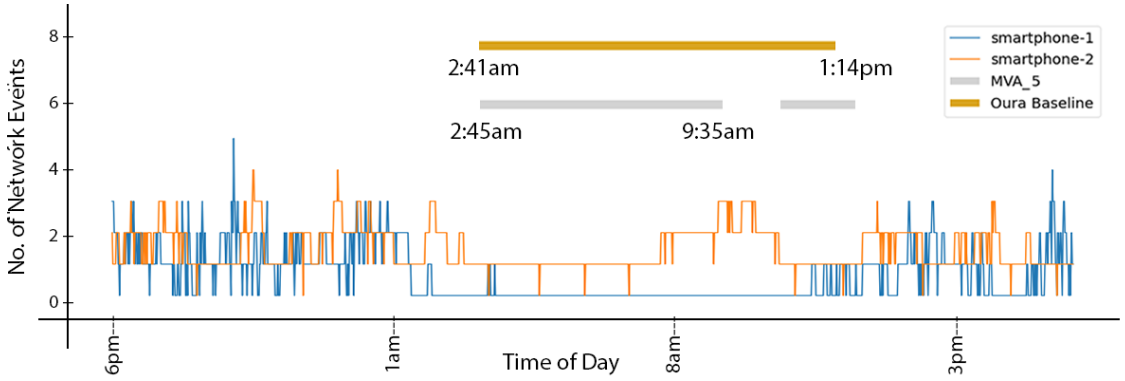

Fig. 20. Considering the most extended sleep period perversely overlooks split-like sleep behavior for P9.

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.

6.2.3 Wake Events Amid Nocturnal Sleep. While our system is modeled to detect nocturnal sleep, *SleepMore* can handle the detection of wake states at per minute granularity from disruption of WiFi events occurring amid long sleep periods. As per Figure 20 for P9, detecting continuous wake states occurring over an hour negatively affected *SleepMore*'s ability to estimate sleep duration that is more accurate to the ground truth. Instead, *SleepMore* registered their wake time to be 9:35 am, despite a continuing sleep period between 11:00 am to 1:00 pm. Considering the longest sleep period can generally help improve overall performance (see Figures 10 and 18 for P2 and P24), but it can perversely overlook other sleep period that comes soon after the wake period. Note, as we explained in Section 4, our sleep baseline only considers nocturnal sleep (meaning a split-like sleep behavior would be consider these split events as a whole).

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. The system is robust to regular and irregular sleepers. However, it does not handle unexpected disruptive sleep events from regular sleepers as well, primarily since it currently lacks samples of users sleeping at dawn. Further, considering the most extended sleep period can impact our model performance, especially among users who exhibit split-like sleep behavior.

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. In what follows, we tested *SleepMore* in a cross-environment. That is, the semi-personalized model for each home participant is built using students' WiFi sleep dataset (the primary study) but with 10% of their data.

Figure 21 plots the classification of sleep and awake states using their semi-personalized random forest model for $homeUser_1$. 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 12, 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 between 2 to 8 minutes of sleep error and 1 to 28 minutes of

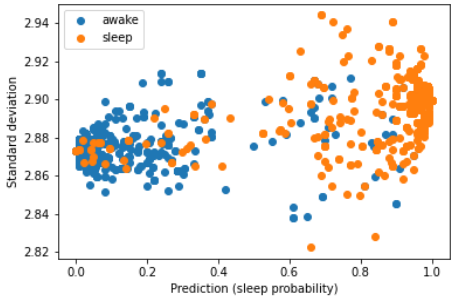

Fig. 21. Classification of sleep states and measure of dispersion for $HomeUser_1$.

	$HomeUser_1$	$HomeUser_2$	$HomeUser_3$
Accuracy	0.986	0.988	0.991
Precision	0.977	0.983	0.986
Recall	0.975	0.978	0.982
F1	0.976	0.980	0.984
T_{sleep} **(<=1%)**	6-8 mins (mean: 6 mins)	4-11 mins (mean: 6 mins)	1-17 mins (mean: 6 mins)
T_{wake} **(<=1%)**	1-28 mins (mean: 8 mins)	1-27 mins (mean: 7 mins)	1- 4 mins (mean: 2 mins)
T_{sleep} **(<=5%)**	2-5 mins (mean: 3 mins)	24 mins (1 sample)	-
T_{wake} **(<=5%)**	3-19 mins (mean: 8 mins)	6 mins (1 sample)	-

Table 12. *SleepMore*'s performance and errors all within 1% uncertainty rates in a private home setting.

wake error. the evaluation yielded similar results for our home users sharing the same residence. Note that all predictions were estimated within a 3% uncertainty rate; therefore, no outliers were removed.

In a traditional sleep setting, "lights off" signals a person's intention to sleep. However, the intention to sleep to actual sleep time (sleep latency) is one measure that varies across persons. To examine potential discrepancies, we compared the time difference produced by our system with $HomeUser_2$ and $HomeUser_3$'s lights off logs and sleep baseline. As per Figure 22, using 'lights off' as a point of reference for sleep produced between 24-90 minutes time difference for our participants compared to the Fitbit wearable, with the largest differences occurring over the weekend. The time difference between our system and 'lights off' ranges from 14-69 minutes. It should be noted that while our system cannot detect sleep onset latency with coarse-grained information, the time difference is considerably less against the Fitbit compared to lights off.

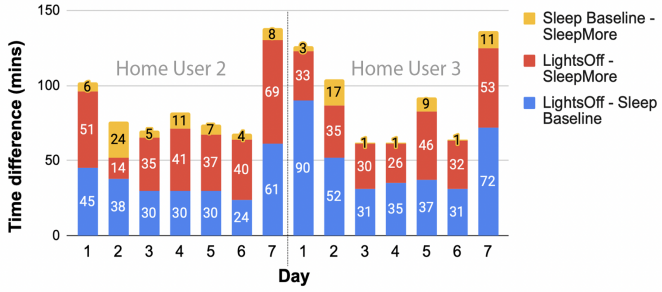

Fig. 22. Time difference of *SleepMore* with lights off and sleep baseline.

	$HomeUser_2$	$HomeUser_3$
Accuracy	0.991 (+1.9%)	0.993 (+0.2%)
Precision	0.991 (+0.8%)	0.989 (+0.3%)
Recall	0.978 (+0%)	0.986 (+0.4%)
F1	0.984 (+0.4%)	0.988 (+0.4%)
T_{sleep} **(<=1%)**	2-7 mins (mean: 4 mins)	5-8 mins (mean: 6 mins)
T_{wake} **(<=1%)**	1-26 mins (mean: 6 mins)	1-8 mins (mean: 4 mins)

Table 13. Performance difference with and without shared device features.

6.3.1 Utility of Smart Home Devices. These results thus far have shown promise in utilizing WiFi data from personal devices. We emphasize personal devices because much work has proven that these device types are a reasonable proxy for estimating user behavior. However, with the increasing use of smart home devices today, we briefly investigate their impact on our model performance. Smart home devices, although shared, are almost

guaranteed to be connected to home WiFi networks. Table 13 presents the performance difference for $HomeUser_2$ and $HomeUser_3$, respectively. In removing features generated by smart home devices, our model performance improves overall accuracy but with insignificant difference (p>.1). Note, however, no one prediction was made beyond 1% uncertainty rate.

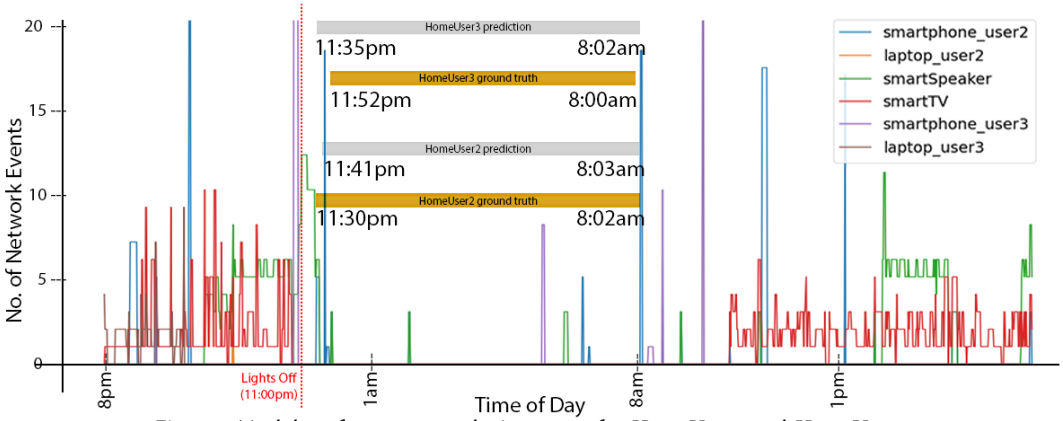

Fig. 23. Model performance on device usage for $HomeUser_2$ and $HomeUser_3$.

It is essential to comprehend the following attributes in considering the relevance of our results. First, for many people, the second-nature action of switching off the TV every night before bed makes it more reasonable to approximate user states. Equally, this behavior may not be true for users who habitually leave their TV on during bedtime. In this case, our users are a couple with similar sleep hygiene who did not exhibit a pattern of leaving their devices running through the night. Second, they displayed similar sleep schedules, with one person shortly sleeping and waking up after another. Monitoring WiFi network data from their smart TV and speaker made little impact on our model performance, whether at night or in the morning. Where this condition does not hold, monitoring WiFi data from shared devices will be less straightforward for sleep monitoring.

One way to accommodate the nuances in users' digital device practices is by distinguishing each device as primary, secondary, or shared use, assigning weights based on utilization frequency and time of day. By operationalizing feature extraction in such a way, we could more accurately determine the user utilizing a shared device. For example, Figure 23 charts the combined use of devices for $HomeUser_2$ and $HomeUser_3$. While both users shared the same device practice through the night, it is unclear who was operating the smartTV, and smart speaker from 10 am onwards. For *SleepMore* to extend its capability in estimating daytime naps, detecting who is awake while utilizing the device during the day will be crucial to our model development.

Key Takeaway The primary objective of this secondary study is to demonstrate the applicability of our system in a private home setting, albeit on a different time scale and different sleep baseline. While the insights we derived were based on a limited observation window, we show how our technique is adaptable to different residential settings as long as users connect their devices to a dedicated WiFi network when at home. In a shared setting, 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 analyses compare *SleepMore*'s performance against a system implementation of three unsupervised learning techniques by prior work [28]. Similarly, we evaluate the performance of these models using leave-one-out cross validation.

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

Mammen *et al.* 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), looking at the rate of change of values in a single variable time-series data; normal prior [14], hierarchical prior [11], and an ensemble of normal, uniform, and hierarchical priors [28]. The authors found that 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.

Building on these findings, we employed the system developed by Mammen *et al.*[28] on our primary study dataset. Table 14 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 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) and sleep onset latency. Currently, only estimation of bedtime and wake time is possible with *SleepMore*. 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. The operational challenges faced in our user study (see Section 4) hindered our progress in experimenting with predicting nap times (note: today, the software update in wearable sleep trackers are equipped with nap detection). With *SleepMore*'s capability to detect awake and sleep states at per minute granularity, the system could conceivably detect different sleep patterns, including polyphasic and split-like sleep behavior. Our estimation technique, at present, smooths out wake events amid a sleep period at a fixed threshold, thus requires implementing dynamic thresholding in setting naps and split-like sleep behavior for different users.

7.2 Extending to Different User Populations and Device Types

The main population target for this study is students and a smaller scale of full-time working professionals with regular work schedules. However, different populations might exhibit different sleep patterns. While students have the flexibility to adjust their sleep schedule to academic/social demands by sleeping during the day, most daytime workers have to follow nighttime sleep. Sleep patterns might be highly irregular in shift workers or patients with insomnia, which our model development has not accounted for in this phase.

Nonetheless, the cross-environment model development of on-campus student residents to home participants exemplifies the feasibility of supporting large-scale sleep monitoring in family dwelling units. Although the supplementary home study was a minor scale, it should be noted that our home participants were not in a single-living 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. We emphasize 'personal devices' because these devices (in our study and similar prior work) have shown to effectively approximate user presence and activity.

We learned from our home study the promise of monitoring WiFi data generated from shared information appliances such as smart TV and smart speakers, which are almost guaranteed to be connected to home WiFi networks. However, the practicality of using these devices grows with increasing complexity when more people share the device. It will also pose as an issue when extending *SleepMore* to detect daytime naps, requiring the model to learn the user more likely to be using the device based on their routines. This is where the personalization of device feature weights may be relevant.

7.3 Improvements to Handle Corner Cases

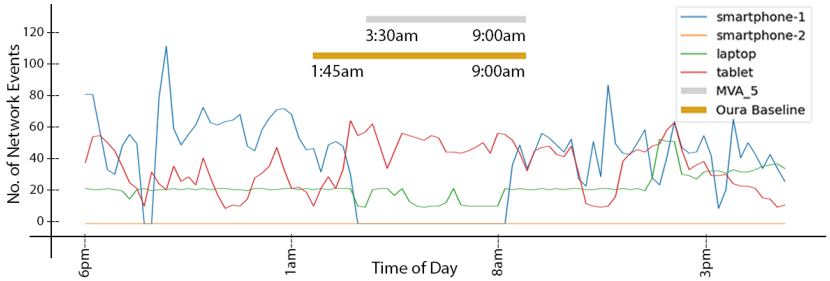

Fig. 24. Outlier with large error.

Finally, Figure 24 exemplifies an error our system encounters that could be attributed to multiple variables and conditions. In fact, it is common for devices to stay connected to the WiFi for activities such as software updates, music/video streaming during the night. As previously suggested, one possibility to reduce the impact of these device features in sleep prediction is assigning different weights to each device as a separate feature.

Further, under conditions where *SleepMore* is utilized as a complementary tool over long-term observations, a feasible way to correct sleep predictions with large errors is by supplementing historical data from another sleep tracker modality to train a user's semi-personalized model continuously. This is also where our work continues with the LSTM implementation, briefly presented in Section 5.1.4. Months of training data will allow us to experiment with building a classifier that compares sleep prediction the following day to longer historical sequences. Instead of using WiFi network event features extracted per day, we can couple our model input with WiFi network data of users over weekdays, weekends, and calendar events, tailored to their device habits and sleep routines. Current work relied on the trial-and-error method for model architecture and hyperparameters. Extension to this effort can include employing more sophisticated optimization approaches for better model configuration.

8 CONCLUSIONS AND FUTURE WORK

This work proposed *SleepMore* as a promising, low cost and easily scalable supplementary solution to commercial sleep trackers for long-term sleep monitoring in residences such as dormitories and homes. *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 used data collected from 46 participants living in on-campus dormitories, predicting 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. We also demonstrated the application of *SleepMore* in private home settings and provide insights into the utilization of shared home devices. A detailed comparative analysis highlighted the importance of multiple user devices to accurately estimate sleep duration 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] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] 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.

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

[4] 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.

[5] 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.

[6] Phuthipong Bovornkeeratiroj, John Wamburu, David Irwin, and Prashant Shenoy. 2021. VPeak: exploiting volunteer energy resources for flexible peak shaving. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 121–130.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] Massimiliano de Zambotti, Leonardo Rosas, Ian M Colrain, and Fiona C Baker. 2019. The sleep of the ring: comparison of the ŌURA sleep tracker against polysomnography. *Behavioral sleep medicine* 17, 2 (2019), 124–136.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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]

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.
[34] Nikolaus C Netzer, Riccardo A Stoohs, Cordula M Netzer, Kathryn Clark, and Kingman P Strohl. 1999. Using the Berlin Questionnaire to identify patients at risk for the sleep apnea syndrome. *Annals of internal medicine* 131, 7 (1999), 485–491.
[35] 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.
[36] Oura Health Oy. [n.d.]. New Oura Ring Generation 3. https://ouraring.com
[37] 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.
[38] 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).
[39] Pew Research Center. April 7, 2021. Internet/Broadband Fact Sheet. https://www.pewresearch.org/internet/fact-sheet/internet-broadband/#who-has-home-broadband
[40] Pew Research Center. April 7, 2021. Mobile Fact Sheet. https://www.pewresearch.org/internet/fact-sheet/mobile/#ownership-of-other-devices
[41] 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/
[42] 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/
[43] 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/
[44] 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.
[45] Kivan Polimis, Ariel Rokem, and Bryna Hazelton. 2017. Confidence Intervals for Random Forests in Python. *Journal of Open Source Software* 2, 1 (2017).
[46] William H Press and Saul A Teukolsky. 1990. Savitzky-Golay smoothing filters. *Computers in Physics* 4, 6 (1990), 669–672.
[47] 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.
[48] 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.
[49] 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.
[50] Robert A Steer and Aaron T Beck. 1997. Beck Anxiety Inventory. (1997).
[51] Robert A Steer, Gregory K Brown, Aaron T Beck, and William C Sanderson. 2001. Mean Beck Depression Inventory–II scores by severity of major depressive episode. *Psychological reports* 88, 3_suppl (2001), 1075–1076.
[52] John M Taub. 1978. Behavioral and psychophysiological correlates of irregularity in chronic sleep routines. *Biological Psychology* 7, 1-2 (1978), 37–53.
[53] 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.
[54] 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.
[55] Amee Trivedi, Jeremy Gummeson, and Prashant Shenoy. 2020. Empirical characterization of mobility of multi-device internet users. *arXiv preprint arXiv:2003.08512* (2020).
[56] Terumi Umematsu, Akane Sano, Sara Taylor, and Rosalind W Picard. 2019. Improving students' daily life stress forecasting using LSTM neural networks. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 1–4.
[57] 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.
[58] 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.

[59] 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).

[60] 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.

[61] 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.

Revision Summary and Reviewer Response

We thank the reviewers for their constructive comments. We summarize the major changes to our paper as follows, and address the individual reviews accordingly. These changes are colored blue on the revised manuscript. Minor revisions in correcting sentence structures, spelling, and grammar, are done throughout the paper.

- **Revised system motivation:** We revised the motivation for our work in several sections throughout the manuscript. We clarified that our approach is best seen as a supplement, rather than a complete replacement, for current sleep monitoring solutions—for example, when users are without their sleep trackers or do not have a dedicated mobile app installed on their smartphone. In order to properly supplement such capabilities, it is essential that our solution, while using coarse-grained WiFi network activity data, provides statistically comparable performance to wearable sleep tracking devices in terms of data accuracy and consistency.

- **Home user study:** We showed the generality of our approach beyond campus students and campus WiFi networks. We conducted a new week-long user study for two participants living in a shared residence. This study is aimed to further demonstrate *SleepMore* in a home setting. Through this study, we addressed reviewers' comments on the shared residence, comparison with traditional 'lights off' baseline, and monitoring WiFi data from shared devices.

- **Neural networks implementation:** We implemented Long Short-Term Memory (LSTM) model into *SleepMore* and provided the LSTM model performance of our system as part of the main results. Our results show comparable performance between LSTM and the random forest algorithm. While LSTM achieves a 3% increase over RF in terms of recall, it has slightly worse precision and overall accuracy. Accordingly, *SleepMore* supports both LSTM and a conventional random forest model for each user.

- **Error analysis:** We added new figures illustrating how our system performs under conditions where users experience irregular sleep habits. We also provided a distribution breakdown of our system predictions under different uncertainty rates.

Meta-review (1AE) comments

(1) Address the issue of only considering students in the study. How would a more diverse set of people/ scenarios/ multiple people/ multiple devices affect the study?
Response: We addressed this comment in several ways. First, our user study on 'Home-Use Applicability' in Section *6.3* was for a non-student non-campus user. We conducted an additional user study to include two new users (both working professionals) residing together. Using a model trained on the student population (plus 40% of their own data), our revised manuscript presents a similar model performance to the first user. We considered the impact of monitoring WiFi data generated from shared devices, primarily the smart TV and smart speakers, and provided insights into the conditions under which our system can use smart home devices for sleep monitoring. These findings supplement our investigation on the impact of sleep prediction with a varying number of personal devices described in Section *6.1*. Finally, we expanded on our discussion to account for a more diverse set of users in Section *7*.

(2) Address the issue of baselines, for example how does the approach compare to a much simpler approach: measuring lights switched in the home, how exactly was the baseline devised (justifying removing some data from Oura's baseline).

Response : In addressing this comment, our new home user study included participants providing additional self-reports of when lights were switched off and on, sleep and wake times, and their Fitbit sleep logs. Our previous user study did not consider this baseline, and we have added it to our new user study as suggested by the reviewers. With the new information supplied, we drew a comparison in Section *6.3* on the time difference in estimating sleep time between (1) lights-off and wearable, (2) lights-off and *SleepMore*, and (3) wearable and *SleepMore*. Secondly, we elaborated on the process our sleep researchers took to devise sleep logs generated from wearable as the baseline in Section *4.2*.

(3) Consider more advanced learning algorithms, such as neural networks.
Response: Our revised manuscript now includes the implementation of LSTM in Section *3.3* and performance for sleep prediction in Section *5.1*. Our results indicate that the performance of deep learning-based LSTM is comparable to the conventional random forest approach – achieving a 3% increase in recall but decreased precision and overall accuracy. Given these comparable results, we report random forest results in subsequent experiments since LSTM results are very similar. We note that *SleepMore* now supports both models (LSTM and random forest) and defaults to the random forest since it is less computationally expensive than LSTM.

(4) Consider the robustness of the study.
Response: To demonstrate the robustness of this study, we expanded our error analysis (originally in Section *6.1*). In this revised manuscript, our findings are reorganized in Section *6.2* to better explain the following:
   (a) Figures 16 and 17 are meant to show how *SleepMore* can maintain its accurate performance in predicting sleep for regular sleepers and irregular sleepers. These cases, however, differ for a regular sleeper with sudden disruptive sleep events.
   (b) Accordingly, as per Figures 18 and 19, we included the performance of *SleepMore* in detecting disruptive sleep for a regular sleeper. We explained why our system produces large errors in these cases, mainly due to the lack of training data.
   (c) We supplemented Figure 20 to demonstrate how *SleepMore* can technically detect a wakeup period amid a nocturnal sleep (i.e., disrupted night sleep). We also note however that extending our approach to true polyphasic sleep (i.e., multiple sleep periods in a 24 hour period) is part of future work. We expanded our discussion to account for other sleep characteristics, including detecting nap time in Section *7.1*.

Reviewer 2 (2AE)

(1) The paper should be further improved by including following items into a discussion. The target users in the paper are college students who can be considered as digital natives. What would be the effects/alternatives/limitations for users being less WiFi driven in their daily behavior.
Response: Our primary user study was among college students, considered digital natives. However, the three home users were full-time working professionals between the ages of 35-46. In Section *6.3*, we clarified that the semi-personalized prediction models for our home users were trained on the WiFi data collected among our student population plus 40% of their data. Figure 23 illustrates the WiFi events generated from the home users' devices, notably more sparse than the student users. Findings from this investigation demonstrated that our technique is scalable to different user profiles with different WiFi usage patterns. However, it includes the prerequisite use of WiFi through personal devices.

(2) How about users taking naps during the day, how could that be accounted for?

Response: Currently, our approach is geared towards detecting the longest inactive period over a 24-hour window as the sleep period and needs to be extended to detect multiple inactive periods corresponding to multiple sleep periods, including daytime naps. As noted earlier, we supplemented a new Figure 20 showing how our technique can handle wakeful events since it predicts at per-minute granularity. Building on this observation, we discussed in Section *7* how our model technique could be extended to assess different sleep patterns, including polyphasic sleep and shorter nap periods.

(3) How does the system deal with disruptive and irregular events, e.g. party, vacation, visitors, celebrations, family events? How does the system with night activities disrupting, e.g. reading, sleep disorders, sex, wake times?
Response: Addressed in Meta 1AE (4).

(4) How does the approach compare to a much simpler approach: measuring lights switched in the home?
Response: Addressed in Meta 1AE (2).

(5) How would the system need to be extended in multi-user homes sharing devices, e.g. tablets, TV?
Response: Addressed in Meta 1AE (1).

(6) Overall, the paper has to tone down:
- Wifi can be a great source of information for measuring sleep. However, as listed above, there are many disturbance factors that have to be accommodated by additional efforts.
  Response: In addressing this comment, we have first toned down the language throughout the paper and clarified the position of our system as a complementary tool for long-term monitoring. Second, we have clarified that our data collection of WiFi data from all devices was done fully in the presence of real-world disturbances, such as the phone receiving background notifications, the user receiving/making phone calls, and the phone running software updates in the background. Users were not instructed to limit their WiFi usage in any way during their sleep period. However, as explained in Section *4*, users were also not instructed to provide self-reports of their wakeful activities for privacy reasons. By further supplementing our error analysis listed in Meta 1AE (4), we have shown that the system is robust to different sleep routines and explained why the model performance is affected by a disruptive sleep event for a regular sleeper.

- The paper should rather focus on complementing other systems by Wifi data than promoting wifi as an alternative to devices, e.g. oura ring.
  Response: Addressed. We have now positioned our approach as a complement to sleep trackers.

Reviewer 3

(1) The performance of the model is low. 15-28 mins sleep error and 7-29 mins wake error needs to be improved.
Response: While our performance errors might seem high numerically at first glance, they are comparable to the accuracy obtained using commercial sleep trackers. A stated goal of our paper is to achieve comparable accuracy to sleep trackers, given that commercial wearable sleep trackers themselves have error as shown in prior medical studies. We refer to a recent study by Massar et al. [29], who analyzed the sleep prediction error range produced using commercial wearable trackers, smartphone screen interaction, and EMA and showed similar error ranges as our approach when using sleep trackers.

Further, we must clarify that in predicting sleep duration, our system must account for sleep onset latency, the time they actually fall asleep after bed, which coarse-grained information such as WiFi network data cannot detect. Given this restriction, we have provided quantitative evidence that coarse-grained information can provide sleep duration comparable to wearable trackers at no statistical difference. The system limitation of not determining sleep onset latency is described in Section *7.1*.

(2) It occurred to me that they have only two contributions - the tool and the study.
Response: We have revised the contributions of our work in Section *1*.

(3) The author considers students from dormitories. Chances are high that all of the students maintain the same schedule to attend classes. In that case, there is not much diversity in training and testing data.
Response: While our main study was among student participants, we have explained in Section *6.3* that our system evaluation for home application was among full-time working professionals living in different residential settings. Further, their semi-personalized models were trained using the student dataset, thus cross-environment. Second, our analysis found varied sleeping patterns among the student population, including circadian phase, night-owl behavior, regular and irregular sleepers. We supplemented these findings, as described in Meta 1AE (4).

(4) How did the author differentiate between multiple users using the WiFi in indoor settings? Family members can use the tablet of other members.
Response: We address this comment in two parts. First, with respect to our participants, the study required them to either provide us with the WiFi MAC addresses (students) or filter out logs that only pertain to their (home) personal devices. We clarified this procedure in Section *4.1*. Second, our new home participants provided us with WiFi logs of shared devices (i.e., a smartTV and a Google smart speaker). Our new findings in Section *6.3*. suggest that WiFi network activity data generated from shared devices remain practical for our couple-users. However, we explained the conditions under which such utilization remains relevant. In Section *7.2*, we discussed the operationalization of *SleepMore* in a multiple-user home setting and how our work continues with a deeper investigation into the nuances of using smart home devices in different family settings.

(5) How did the author plan to handle the cases when users have multiple devices connected to the same WiFi and running application in the background? For example, during an incoming phone call in messenger. At that time, all the tablets, and phones will use the WiFi and the tool will detect the wake time for the user.
Response: As responded in R2 (6), the data collection of WiFi data from all devices already included such everyday disturbances, including the phone receiving background notifications, the user receiving/making phone calls, and the phone running software updates in the background. Indeed, our analysis found the negative impact of noisy WiFi data in yielding large prediction errors. It is for this reason that we implemented uncertainty quantification, in addition to employing an ML-based mechanism to handle such cases. Further, we have supplemented new graphs, also clarified in Meta 1AE (4), illustrating our system's ability to detect wake events.

(6) What type of devices did the author consider for the study? Did they consider smart home devices like-Robo vacuum, smart TV, and Smart Oven. Different people may use the smart speakers for bedtime stories or for meditation before sleep. How did the author plan to extract those characteristics?
Response: We addressed this comment in several ways. First, we provided details about the type of devices users had consented to us collecting their WiFi network event data as per Tables 1 and 2 in Section *4*.

Second, we included a new home study, allowing us to collect WiFi network event data generated from the smartTV and smart speaker. In Section *6.3*, we explained the conditions under which shared smart devices can be used to estimate users' sleep. However, as noted by the reviewer, different users will have different utility habits. We discussed our continuing effort toward this complexity in Section *7.2*.

(7) What is the expertise of the researchers to filter out incorrect data from Oura's baseline? And why the author didn't consider Oura baseline for a private residence setting?
Response: We clarified in Sections *1* and *4* that the user study protocol was designed and run by qualified biomedical scientists/experts in a sleep medicine laboratory. They, too, were in charge of processing ground truth information, comparing sleep logs generated from wearable devices and EMA used in developing our system. We added new home participants for our revision who preferred logging sleep through their personal Fitbit device.

(8) It would be interesting to compare the performance of commercially used sleep tracker devices with the author.
Response: Yes, such work has been published as we responded earlier in R3 (1) – a group of sleep medical researchers, Massar et al. compared commercial sleep trackers with smartphone screen interaction and self-reported EMA [29]. Our proposed technique aims to address the practical challenges that sleep researchers alike face in conducting longitudinal studies.

Reviewer 4

(1) The authors claimed "Users who have the most disrupted sleep experiences tend to be those who do not wear sleep trackers" - This appears to be intuitive. But do they have a supporting citation for the claim?
Response: As part of our revision, this statement has been removed from Section *1*.

(2) The authors claimed that "It is essential to underscore that our study defined no inclusion/exclusion criteria." But I am wondering how would "having multiple devices" be a condition. It seems like the experiment would not "require" participants to have more than one device. However, this is the main contribution of the paper, so I would suggest the authors' tone down this part a bit.
Response: We intended to explain that our study made no distinction in considering users with different sleep profiles. Indeed, participants were required to supply the MAC addresses of their multiple devices. We have corrected this phrasing in Sections *1* and *4* of our revised manuscript.

(3) Related, in Table 5, there are only 28% of the user have a study smartphone, while the study procedure stated that all participants would get a study phone. Please explain this.
Response: We have corrected this statement and clarified in Section *4* that participants were issued a study-smartphone Android device if they owned an iPhone; Android users were allowed to use their normal phone instead of getting a separate study phone. The study-smartphone came preinstalled with an Android-based logging app for EMA.

(4) After the fine-grained prediction, the authors decided to use the longest sequence of sleep states, without a clear justification. Why not discard the edges with low confidence?
Response: In Section *5.3*, we explained why discarding the edges with low confidence is not a straightforward solution to estimating sleep duration. We supplemented Figure 10 as a typical example, informing our

decision to explore other techniques such as smoothing.

(5) Related to this question, the authors proposed to remove the days with low confidence and observe a performance increase (Fig. 8). This is like "throw the bad apples out of the basket". The more important is how much data is left. The authors mentioned some numbers in the text (10% with the 1% threshold), but I would recommend putting them in some figures, as this has important deployment implications.
Response: We address the comment by including a new Figure 12, showing the breakdown of our predictions at different uncertainty rates. With most predictions ranging between 0.02-0.05, we excluded 95 of 490 outcomes. Restricting the threshold to 1% retains 12% of these outcomes.

(6) I am also curious to know the authors' discussion on how to further deal with these days with low confidence besides "ignoring" them.
Response: Indeed, our work continues to explore other sophisticated methods besides ignoring the bad predictions, albeit a small amount. In Section *7.3*, we explained how our current implementation of LSTM remains promising as we investigate the impact of model performance with learning longer historical sequences.

(7) In the ML results part, it is unclear what kind of validation process was used. Is that a leave-one-user-out? Or a user-level train/test split? Related, how did the authors tune the model hyperparameter? On the training set or testing set? For the semi-personalized model, what are the 40%? The first 40% of the data? Or randomly 40%? Please clarify these details.
Response: We have clarified in Section *5* that our model evaluation is based on a leave-one(user)-out validation. For each user, we performed a 5-fold cross validation to randomly sample 40% of their data for semi-personalization.

(8) There are some details missing in Sec 6. In Sec 6.3, it is unclear what the model setup is. Please add more details. In Sec 6.4, did the authors re-implement these techniques? If so, please clarify. If not, please state the source of these results.
Response: Our models for home users were tested in a cross-environment. We clarified in Section *6.3* that for the home users, each model was built with students' WiFi/sleep data plus 40% home user's data. We also revised the writing in Section *6.4* that we obtained the code from the prior work authors and employed their system to train and test on our dataset.

Reviewer 5

(1) It is standard to compare various types of machine learning algorithms, but the work does not include more state-of-the-art neural-network-based models. Even an MLP would be quick to train but could provide insights and improve the impact of the work. Is there a reason this common model choice was not validated?
Response: Addressed in Meta 1AE (3).

(2) Additional details regarding how the ground truth was obtained would be helpful, given the minimalistic study.
Response: We supplemented details of our user study, including user demographics, screening tools, and procedure in Section *4* of our revised manuscript. We also clarified that our study was designed and conducted by researchers in the sleep medicine laboratory, trained to validate ground truth data from sleep

trackers and EMA.

(3) Given that most of the study was done within one type of environment, it is not clear how robust the noise model and parameter settings are to other environments and device types.
Response: Addressed in Meta 1AE (1) and (4).

(4) Given the emphasis on scalability, do the authors expect their model and findings to generalize across environments? Can the model be tested in cross-environment (i.e., trained on dorms, tested on the single home participant), to gain insights into the impact of such ambient/platform settings and need for potential calibration of the current extensively tuned SleepMore model?
Response: Addressed in R4 (8).

(5) Minor comment: Eqn. for C_i on page 9 has an error.
Response: Fixed.

(6) Will the dataset be made public, or is that not possible? This could increase the significance of the work.
Response: Currently, the terms of the IRB procedure do not allow the release of WiFi data since it contains privacy-sensitive location information of users. To improve the usefulness of our work, we are discussing the release of some partial information, including home user study data. The code for all the work will be made open source when preparing the camera-ready paper.