

An Active Sleep Monitoring Framework Using Wearables

H M SAJJAD HOSSAIN, SREENIVASAN R. RAMAMURTHY, MD ABDULLAH AL HAFIZ KHAN, and NIRMALYA ROY*, University of Maryland Baltimore County

Sleep is the most important aspect of healthy and active living. Right amount of sleep at the right time helps an individual to protect his physical, mental, cognitive health and maintain his quality of life. The most durative of the Activities of Daily Living (ADL), sleep, has a major synergic influence on a person's functional, behavioral and cognitive health. A deep understanding of sleep behavior and its relationship with its physiological signals, and contexts (such as eye or body movements) is necessary to design and develop a robust intelligent sleep monitoring system. In this paper, we propose an intelligent algorithm to detect the microscopic states of the *sleep*, which fundamentally constitute the components of a good and bad sleeping behavior and thus help shaping the formative assessment of sleep quality. Our initial analysis includes the investigation of several classification techniques to identify and correlate the relationship of microscopic sleep states with the overall sleep behavior. Subsequently, we also propose an online algorithm based on change point detection to process and classify the microscopic sleep states. We also develop a lightweight version of the proposed algorithm for real-time sleep monitoring, recognition and assessment at scale. For a larger deployment of our proposed model across a community of individuals, we propose an active learning based methodology to reduce the effort of ground truth data collection and labeling. Finally, we evaluate the performance of our proposed algorithms on real data traces, and demonstrate the efficacy of our models for detecting and assessing the fine-grained sleep states beyond an individual.

CCS Concepts: • **Information systems** → **Information systems applications; Mobile information processing systems; • Computing methodologies** → **Machine learning; Learning settings; Active learning settings; • Computer systems organization** → **Embedded and cyber-physical systems; • Hardware** → *Communication hardware, interfaces and storage; Sensor applications and deployments;*

Additional Key Words and Phrases: sleep monitoring, active learning, crowdsourcing, gradient classifier, wearable technology

ACM Reference Format:

H M Sajjad Hossain, Sreenivasan R. Ramamurthy, Md Abdullah Al Hafiz Khan, and Nirmalya Roy. 2017. An Active Sleep Monitoring Framework Using Wearables. *ACM Trans. Interact. Intell. Syst.* 0, 0, Article 1 (May 2017), 30 pages. <https://doi.org/0000001.0000001>

*The corresponding author

This work is supported partially by the UMB-UMBC Research and Innovation Partnership Grant, NSF Award #1344990, and Constellation E²: Energy to Educate Grant.

Author's addresses: Nirmalya Roy, Department of Information Systems, University Maryland Baltimore County, Baltimore, 21250, US..

Authors' address: H M Sajjad Hossain, riaj.sajjad@umbc.edu; Sreenivasan R. Ramamurthy, rsreeni1@umbc.edu; Md Abdullah Al Hafiz Khan, mdkhan1@umbc.edu; Nirmalya Roy, nroy@umbc.edu, University of Maryland Baltimore County.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2017 Association for Computing Machinery.

2160-6455/2017/5-ART1 \$15.00

<https://doi.org/0000001.0000001>

1 INTRODUCTION

Sleep monitoring has been in the limelight of research due to the growing need of good quality sleep in person's day to day lives. As well as sleep has a bi-directional relationship with the well-being of a person [Saeb et al. 2017]. Sleep disorders such as sleep apnea, COPD (Chronic Obstructive Pulmonary Disease), Chronic renal diseases and various other medical conditions can manifest itself as disruptions in sleep patterns and thus affect sleep quality [Parish 2009]. [Kline 2013] has defined sleep quality as one's satisfaction of the sleep experience, integrating aspects of sleep initiation, sleep maintenance, sleep quantity, and refreshment upon awakening. Sleep quality is highly correlated to exhaustion, discomfort, depression and lack of concentration during the day. The quality of sleep can affect the intuitive symptoms caused by the underlying disease. Clinical studies have suggested that if the sleep quality is improved, the underlying symptoms of a patient might improve too. Monitoring sleep could help a physician to diagnose the underlying condition. Clinically, Polysomnography (PSG) is used for sleep monitoring, which is also the "gold standard" for sleep monitoring and diagnosis of sleep related disorders. PSG captures multiple physiological parameters such as Electroencephalogram (EEG), Electrocardiogram (ECG), Electromyogram (EMG), Electrooculogram (EOG) simultaneously which makes it the best solution to diagnose sleep disorders. PSG provides general sleep measures, such as total sleep time (TST) and sleep efficiency (SE), and also detects specific sleep stages. Conducting PSG requires specialized environment like sleep center or clinic which makes it un-suitable for day to day sleep monitoring. Recent advances in the field of sensor technology and wearables have led to the advent of various sleep monitoring wearable devices such as [Fitbit 2007], Actiwatch [Actigraph 2004], BASIS watch [Basis Band B1 2014], Misfit Shine, Withings Pulse O2, etc. These devices are commercially available in the market except for the BASIS watch [Basis Band B1 2014]. These devices record accelerometer data, heart rate to monitor sleep quality. Research indicates that Actiwatch and Fitbit captures the TST and the SE well. In other words, it focuses on calculating the duration of sleep which gives a very good insight on someone's sleep hygiene. However, the BASIS B1 band [Basis Band B1 2014] classifies sleep stages (REM, light sleep, deep sleep) and provides the means to identify patterns and triggers which are causing sleep disturbances [Mantua et al. 2016]. The use of wearable devices has also been proved effective for in-home sleep measurements and evaluations [Kuo et al. 2017].

To evaluate a person's sleep quality, it is important to detect the stages and micro stages of sleep. Researchers view this problem as a classification problem and have been using machine learning algorithms on the data extracted from the wearable devices. Existing sleep monitoring studies have been using supervised learning algorithms where they collect and label a set of training data with pre-defined classes which the system aims to detect. Larger labeled training dataset with consistent label information can help to build a more generalized classifier [Liao and Zhu 2014] [Sordo and Zeng 2005]. One of the major challenges faced by researchers while addressing this problem is the collection of the ground truth information. Collecting ground truth without violating privacy (using cameras) of the individuals is an extremely difficult task. It is also a herculean task for the test subjects and other human annotators to annotate the data manually. Leveraging unsupervised algorithm can help to eliminate the requirement for labeled data. However drawing boundaries between similar instances (with respect to properties) but belonging to different classes are difficult using unsupervised learning. The microscopic sleep states are very hard to differentiate based on the accelerometer data as the signal patterns are quite similar for different states. For example, subtle movements during sleeping and not sleeping while lying in the bed exhibits similar signature pattern. Feedbacks from the users can help us acquire proper labels of data instances for similar circumstances and help us to differentiate between them. By employing Active learning (AL) [Settles 2012], we can acquire feedbacks from the users for important data instances. AL

can not only mitigate the manual effort needed for collecting ground truth information but also reduce the training time. AL retrieves the most informative data instances from a pool of unlabeled data instances and pose them as queries to the annotators. As a consequence we only have to label a handful number of instances. Several researches showed the effectiveness of AL in activity recognition domain [Bagaveyev and Cook 2014] [Hossain et al. 2017] [Alemdar et al. 2011]. Using AL we can improve our model incrementally based on the feedback provided by the annotators. AL works in an online manner, where we receive a stream of instances and then form a pool from which we select the single most informative instance. In order to get bulk label information Crowdsourcing [Chittilappilly et al. 2016] has been exploited in different problem domains. It has been used in activity recognition domain as well [Chang et al. 2017] [Hwang and Lee 2012]. We can collect ground truth information of unlabeled samples in a bulk which will help to improve our model. In addition, identifying and classifying only pre-defined sleep stages is not sufficient for medical diagnosis. Even though the current state of the art wearable technology cannot replace the PSG in clinical evaluations, obtaining accurate sleep stages has been the aim of every researcher who works on this problem domain. Further, we have identified certain weaknesses in the literature such as patients suffering from nightmare disorder, muscle contractions have not been considered in any study. In our previous work Sleep Well [Hossain et al. 2015], we proposed a sleep monitoring model using accelerometer mounted wearable device to classify previously unseen various sleep states, that further improves patients sleep hygiene by being able to pinpoint the causes of sleep disturbances. We trained our model using supervised and unsupervised learning algorithms and identify the basic sleep states - Rapid Eye Movement (REM), non-REM (NREM) sleep, awake, movement, getting up from bed, getting up and sitting. In this paper, we extend our work and introduce a crowdsourcing model to collect large volume of labels and also introduce a sleep scoring module.

Wearable devices have become common and accessible to people these days and researchers are making them evolve by incorporating various sensors as attested by the new release of smart watches such as Google Android Wear [Android Wear 2014], etc. The world is moving more towards wearable technology which has expedited a plethora of application domains ranging from health care applications [Jonas et al. 2014] to sports [Daiber and Kosmalla 2017]. In this study, we have used two different wearable devices - EZ430-Chronos [EZ430-Chronos 2013] and wActiSleep BT [Actigraph 2004] worn on the waist. After collecting the data, we apply a variant of gradient descent algorithm to build a classification model. Further, we apply importance weighted active learning to label the uncertain data points and also incorporate previously unseen sleep states. Active learning improves the annotation effort greatly and improves the performance of classification model. We also exploit crowdsourcing to collect bulk ground truth information in offline. In order to discover abrupt changes on the data streams, increase the classification accuracy, remove noises and provide greater support for informativeness in active learning, we propose an online change point detection algorithm. Finally, we show the results of our proposed algorithm using publicly available benchmark dataset [Borazio et al. 2014] which provides the sleep phases determined by clinical polysomnography where the data was collected using wrist worn device. The main contributions of the paper are summarized below:

- We investigate several classification approaches and propose a gradient descent classification model for recognizing the underlying microscopic context states associated with the sleep disorders.
- We introduce an online change point detection based classification approach to detect any abrupt changes on streaming dataset for better microscopic sleep state classification, data noise and uncertainty reduction.

- We develop an active learning based sleep monitoring model which extensively reduces the data annotation effort and ground truth data collection from the user personal space and help scale the model among a community of individuals.
- We demonstrate a crowdsourcing model using visual illustration and rank the workers based on their feedback.
- We evaluate our model based on two real-world datasets, one with polysomnography result along with the wrist worn accelerometer sensor values from 42 subjects [Borazio et al. 2014], other with general labeled data collected from 17 test subjects.
- We evaluate the quality of sleep using PSQI [Buysse et al. 1989] and Webster [Webster et al. 1982] scale.

2 RELATED WORKS

In medical studies Polysomnography (PSG) is the major sleep study to diagnose a patient's sleep quality [et al 1997]. Polysomnography records biophysiological changes that occur during the sleep. Apart from PSG, some other sleep studies are Multiple Sleep Latency Test (MSLT) and Maintenance of Wakefulness Test (MWT) [Johns 2000]. These diagnoses are cumbersome and need a lot of prior setup, for example, in case of PSG it requires 12 channels requiring almost 22 wire attachments to a patient. Obviously this imposes a great level of discomfort to the patients and its users. Early literatures [NR 1997] involving wearable devices to replicate the Polysomnography results validated the applicability of actigraphy in sleep monitoring. [Sadeh 2011] provided further justification for the use of actigraphy in sleep research. Van et al. proposed a model to support the feasibility of a continuous home-monitoring of sleeping trends using wearable devices [Van Laerhoven et al. 2008]. Recently the authors of [Kuo et al. 2017] proposed a wearable actigraphy device with low sampling rate for in-home sleep assessment. Several other literatures [Sathyanarayana et al. 2017] [Purta et al. 2016] [Sun et al. 2017] [Matsui et al. 2017] [Sathyanarayana et al. 2016] also ascertain the strength and simplicity of wearable devices in sleep monitoring. The authors of [Rofouei et al. 2011] developed a wearable neck cuff system for monitoring physiological signals in real-time. The authors of [Nguyen et al. 2016] have developed a light-weight and inexpensive in-ear wearable sensing system that can capture electrical activities of the brain, eye and facial muscles. [Nguyen et al. 2016] have used a supervised non negative matrix factorization algorithm to adaptively analyze the signals. A sleep monitoring model using image analysis has been proposed in [Nakajima et al. 2000], but it has proved inefficient in case of low light condition at night. [Liao and Yang 2008] used near-infrared cameras to overcome this challenge but the images still created non-uniformity. A novel sleep monitoring framework- LullaBy to capture and monitor the sleeping environment using microphone, light sensor and motion sensor has been proposed in [Kay et al. 2012]. Yanzhi et al. [Ren et al. 2015] put emphasis on the importance of breathing pattern while sleeping and the proposed model captures the breathing sound using signal envelop detection on the acoustic data. The proposed model can detect snore, cough, turn over and get up using the acoustic features. The authors of [Zhang et al. 2013] proposed a real-time system to monitor the sleep conditions where pulse oximeter is exploited to monitor user's pulse oxygen saturation (SPO2) during the sleep process.

Pressure bed sensors have been used to supervise the postures and movements of the users in sleep [Foubert et al. 2012] [Nam et al. 2016a]. Though these methods are unobtrusive and do not create discomfort to the users, but still it has not been streamlined due to its cost and deployment issues. [Hoque and Stankovic 2010] used fine-grained body positions from accelerometer data using WISP tags attached to the sides of a bed. A novel framework for pressure image analysis to monitor sleep postures including a set of geometrical features for sleep posture characterization and three sparse classifiers for posture recognition has been proposed in [Liu et al. 2013]. The authors of

[Nam et al. 2016b] have proposed a sleep monitoring framework comprising of an accelerometer and a pressure sensor. Features pertaining to body motion, respiration, body activity and heart rate were extracted and the proposed framework fuses information from various features and detects the stages of sleep. [Odunmbaku et al. 2016] have proposed a combined framework for fall and sleep monitoring of elderly people by hypothesizing that the acceleration calculated from the accelerometer data will be in the range $0 - 1.5m/s^2$. The authors of [Velicu et al. 2016] have used a custom made accelerometer chip which streams data to an Arduino board. In addition to the accelerometer, a ECG sensor is also used. Features such as Heart Rate and RR interval were extracted and Kushida's algorithm-derived equation was used to differentiate the sleep stages.

Sleep related research are gaining attention due to the recent proliferation of low-cost easy-to-deploy technologies based on mobile and ambient sensors and its large penetration in the market. Commercial wearable devices, such as Fitbit [Fitbit 2007], Zeo [Zeo 2003], Actigraph [Actigraph 2004], Jawbone, Sleep Tracker etc have been used extensively these days for monitoring sleep and activities of daily living (ADLs). iSleep [Hao et al. 2013] uses the built in microphone sensor of smartphone to detect the events which are closely related to sleep quality like body movements, coughing, snoring etc. The authors of [Fahim et al. 2013] used the accelerometer sensor of the smartphone to track the sleep duration and user movement patterns. [Chen et al. 2013] proposed a passive approach to track some stationary features, such as user silence, ambient light, phone usage and charging etc for monitoring sleep habits and developed a mobile application BeWell [Lane et al. 2011] for unified health monitoring. [Bai et al. 2012] used the daily context information of an user to define the sleep quality. Sleep Hunter [Gu et al. 2014] used the accelerometer and microphone sensors of the smartphone a fine-grained detection of sleep stage transition for sleep quality monitoring and intelligent wake-up call. Mimo Baby Monitor is a bodysuit for infants aged 0 - 12 and incorporates respiratory sensor, accelerometer sensor, and temperature sensor to measure the physiological signals, body movements, and temperature respectively. These signals are transmitted via bluetooth to an online data cloud and to the caretaker's mobile [Mimo 2016]. Using several android apps like *Sleep As Android*, *Sleep Time Smart Alarm Clock* [Sleep Time Smart Alarm Clock 2015], *Sleep cycle*, *SleepBot*, etc. it is possible to monitor the quality of sleep. Other commercial unobtrusive technologies like Beddit [Beddit 2015], Hello Sense [Hello Sense 2014] can also monitor sleep. Hello Sense also tracks the quality of the sleeping environment. [Kaplan A 2001] proposed to use change-point segmentation on PSG data to differentiate macrostructural organization of sleep. A point process based novel model for the assessment of heart rate variability and respiratory sinus arrhythmia based on PSG data has been proposed in [Citi L 2011].

Active learning has been investigated in activity recognition domain in several literatures [Bagaveyev and Cook 2014] [Hossain et al. 2017] [Alemdar et al. 2011]. Enamul et al. [Hoque and Stankovic 2012] proposed an activity recognition model AALO in single inhabitant smart home context using active learning. The authors of [Hossain et al. 2017] applied active learning by expected error reduction analysis in a smart home environment for classifying activities which include sleep. In [Sahami Shirazi et al. 2013] the authors propose a model for crowdsourcing sleep data. Yung-Ju et al. proposed a mobile crowdsourcing model to annotate travel activities in real world settings [Chang et al. 2017]. [Hwang and Lee 2012] proposed a crowdsourcing framework that models the combination of scene, event, and phone context to map the unseen audio data with activities.

In this paper we take a different approach and look into the fundamental problem of scaling the sleep monitoring models beyond a specific individual. To realize this first we analyze the microscopic physiological contexts and psychological clauses behind a sound or bad sleep. We investigate traditional classification algorithms to successfully detect those events and propose a

novel online change point detection based method for enhancing the classification accuracy and eventually help guide the design of a community scaling model using active learning.

3 OVERVIEW OF SLEEPWELL FRAMEWORK

Sleep is not just a dormant part of our lives, we remain very active and pass through several important stages of sleep. Interference or disturbance in these states can cause impatience, drowsiness and lack of concentration during the regular activities of daily living. Therefore for maintaining a good sleep hygiene we have to sleep a certain amount of time in each of those sleep states. There are two main types of sleep states:

- Non-Rapid Eye Movement (NREM) (also known as *quiet sleep*). NREM consists of three states (stage-1, stage-2, stage-3).
- Rapid Eye Movement (REM) (also known as *active sleep*).

A complete and healthy cycle of sleep consists of a progression from states 1 to 3 before reaching REM state, then the cycle starts over again. If the REM sleep is disrupted and the person wakes up, the person's circadian cycle is disrupted. In order to complete the cycle the person will move to REM state directly next time. Thus it is very important to sleep a good amount of time each day and maintain a good sleep cycle. REM sleep is considered as active sleep because in this state people dream. If a person is having a nightmare disorder too often it is possible that he/she is having problems to complete the sleep cycle. In this paper, we first focus on properly classifying the sleep cycle into these finer states. We also propose to integrate some other broader intermediate sleep states such as *movement*, *getting up from bed* and *getting up and sitting*. These other states would help to identify the casual and formal causes of sleep disturbance and sleep latency and provide meaningful insights on designing scalable sleep monitoring technologies and automated assessment methodologies.

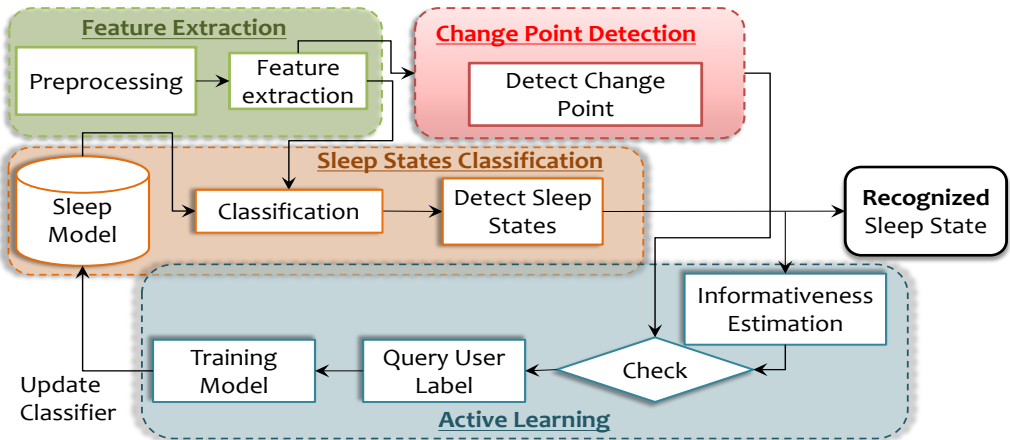


Fig. 1. An architectural overview of Sleep Well [Hossain et al. 2015] Framework

3.1 Sleep Well Architecture

In figure 1 we demonstrate our proposed framework. Our proposed framework consists of the following logical components.

- **Feature Extraction:** After collecting raw sensor data, this component preprocesses and extracts the low level signal features as shown in Table 1 from the processed raw sensor data. (Details in 4.1).

Table 1. Features used for Sleep Micro-States Classification

	Name	Definition
Time Domain Features	Mean	$AVG(\sum x_i), AVG(\sum y_i), AVG(\sum z_i)$
	Mean-magnitude	$AVG \sqrt{x_i^2 + y_i^2 + z_i^2}$
	Magnitude-Mean	$\sqrt{\bar{x}^2 + \bar{y}^2 + \bar{z}^2}$
	Variance	$VAR(\sum x_i), VAR(\sum y_i), VAR(\sum z_i)$
	Co-Variance (Two-axis correlation)	$cov(xy); cov(yz); cov(xz)$
	Standard Deviation	$\sigma_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}; \sigma_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n-1}}; \sigma_z = \sqrt{\frac{\sum(z-\bar{z})^2}{n-1}}$
Frequency Domain Features	FFT-Magnitude	$m_j^{(x)} = a_j + b_j i ; m_j^{(y)} = a_j + b_j i ; m_j^{(z)} = a_j + b_j i $
	FFT-Energy	$\frac{\sum_{j=1}^N (m_j^2)}{N}$ for x, y, z respectively

- Change Point Detection:** After extracting features and analyzing the sleep data, we noticed that change point occurs in sleep transitions (transition from one stage to another, for example unconscious movement during sleeping, waking up, being restless in bed etc.). The importance of these change points has proven to be very effective as it help removing noises in the data and detect the exact point of the sleep transition. For example, when a person gets up from the bed and starts walking, the accelerometer readings other than sleep classes become noisy. Therefore by identifying change points we can partition the data and have more fine-grained information for easing the training effort. (Details in 4.2).
- Classification:** At this stage we train our model using the features from processed raw sensor data and build up our classification model to recognize the several intermediate sleep states. We investigated an online gradient descent [Karampatziakis and Langford 2011] as our classification algorithm. This is different from traditional gradient descent by dealing with importance weights to collaborate with active learning and learning reductions. To calculate the average loss during the classification process we propose to use squared loss function (Details in 4.3).
- Active Learning:** After feeding the test data into our classification model and getting the prediction, active learning helps to calculate the informativeness of each data points. If any data point falls within an uncertain space and while predicting it is found to be the most informative, then if the actual label of the point is provided it would have more significant impact on the classification model. This component then initiate prompt for “*query user label*” and get the ground truth from the user. Subsequently the labels are then used for re-training and updating the model. This component helps to ensure better classification accuracy with minimal user feedback. This also helps to scale the sleep monitoring model across multiple individuals. The input from change point detection method strengthens the active learning query selection by asking the user to label the appropriate sleep state transitioning step. (Details in 4.4).
- Crowdsourcing:** For large scale deployment we apply crowd sourcing for reducing the ground truth collection effort. The problem with existing crowd sourcing platforms is that there is no standard to evaluate the quality of the workers. In our model we calculate two parameters for each worker - *reliability* and *awareness*. Using these two parameters we rank the user and get the most out of our crowdsourcing platform (Details in 4.5).

4 SLEEP WELL FRAMEWORK DESIGN

In this section we describe in details the design of our Sleep Well framework. We first discuss about several micro-states of sleep and feature extraction process. Next we discuss an online change point detection algorithm to have a better handle on the microscopic sleep state classification problem.

4.1 Sleep Event Detection and Feature Extraction

We extract low-level features using each of the three components of the triaxial accelerometer signal to capture the aspects of movements while sleeping. We use both time and frequency domain features in our framework. As the user is not physically active while sleeping, very few number of movements are involved, so we choose a lower sampling frequency. We extract features from data using windows of 60 samples, corresponding to 1 second of accelerometer data. From each window we calculate the features mentioned in Table 1. Time domain features help to differentiate dynamic to static movements. The frequency domain features help identifying patterns within acceleration data, which aids in discriminating discrete movements and their intensities.

4.1.1 Feature Selection. Our model is primarily focused on community scaling, so fewer features will scale the model computationally effective if we can achieve similar accuracy with more features. We select the subset of features, best fit for our model by applying Restricted Forward feature Selection (RFS) algorithm [John et al. 1994]. It was performed in two steps. First we applied Forward feature Selection (FS) algorithm which ranks the features in decreasing order of their accuracy. The FS algorithm iterates through the feature space and measures the Leave-One-Out-Cross-Validation (LOOCV) error for each component in the feature space $\{f_1, f_2, f_3, \dots, f_N\}$. In case of traditional FS, after the first iteration, FS calculates the best individual feature f_i . In the next iterations, FS finds the best subset consisting of two components, f_i and one other from remaining $N - 1$ features. In the following iterations, FS ranks more features and evaluates the subset accordingly, so that after N iterations, the winner is the overall best feature set in these N iterations. In the second step we invoke the RFS to restrict the number of features to rank at each iteration. After the first iteration we consider only the first $N/2$ ranked features for the following iteration. After adding another feature to the winner of the first iteration at the second iteration, we consider the first $N/3$ components of the remaining ranks. RFS repeats this process until it finds the best m feature sets. The difference between conventional FS and RFS is that RFS considers only a part of the remaining ranked features, whereas FS considers all the features. Out of 8 features, the feature selection algorithm chose 4 features (FFT-Magnitude, FFT-Energy, Mean-Magnitude and Co-Variance) which help to attain classification accuracy closer to using all the 8 features.

4.2 Change Point Detection

Change point detection helps find the abrupt variations in the sleep data stream. While some change points provide meaningful insights and some not, our motivation in this work is to find the sleep transitions by calculating the change points (abrupt signal changes) and distinguish between the important and unimportant changes. This is not only helpful to detect the sleep-related events appropriately but also help remove noisy data points from the dataset. We develop a Bayesian online changepoint detection [Adams and MacKay 2007] based algorithm for finer sleep related event identification and online data noise reduction.

We first partition the entire sleep dataset in different regions based on a run length [Adams and MacKay 2007]. Let, $x_{1:N} = \{x_1, x_2, x_3, \dots, x_N\}^T$ denote the N data points observed over time T which is divided into non-overlapping partitions. Consider if we find K change points then let the data set of partitioned data be $\{\rho_1, \rho_2, \rho_3, \dots, \rho_k\}$ at time indices $\{t_1, t_2, t_3, \dots, t_k\}$ where by definition $t_0 = 0$ and $t_{k+1} = N$. The discrete probability distribution over a time interval t_i to t_j is denoted

by $g(t_i - t_j)$. Each partition ρ_t denotes a segment of the data at time t . The length of the each partition or time since the last change point occurred, is defined as “run length”, r . The run length goes back to 0 if change point occurs, otherwise it increases by 1 as follows.

$$r_n = \begin{cases} 0, & \text{if changepoint occurs at } (n - 1) \\ r_{n-1} + 1, & \text{otherwise} \end{cases}$$

The conditional probability that a change point occurs on time t_k after the last change point at time t_{k-1} is

$$P(t_k | t_{k-1}) = g(t_k - t_{k-1}), \text{ where } 0 < k - 1 < n \quad (1)$$

We assume that the predictive distribution of a change point at any time instant t only depends on the recent data. So the change points are assumed to follow markov process. Thus the prior probability of a change point at a time instant t_k is dependent on the probability distribution of the observed data over the time interval and the preceding change point.

$$P(t_k) = \sum_{i=0}^{k-1} g(t_k - t_i) P(t_{k-1}) \quad (2)$$

The change point detection algorithm finds the number of change points and their position by calculating the posterior probability $P(r_n | x_{1:n})$ and integrating it with the predictive distribution $P(x_{n+1} | x_n)$. We do this by calculating the joint distribution of the current run length and observed data $P(r_n, x_{1:n})$.

$$\begin{aligned} P(r_n, x_{1:n}) &= \sum_{r_{n-1}} P(r_n, r_{n-1}, x_{1:n}) \\ &= \sum_{r_{n-1}} P(r_n, x_n | r_{n-1}, x_{1:n-1}) P(r_{n-1}, x_{1:n-1}) \\ &= \sum_{r_{n-1}} P(r_n | r_{n-1}) P(x_n | r_{n-1}, x_{n-r:n}) P(r_{n-1}, x_{1:n-1}) \end{aligned} \quad (3)$$

Where $P(r_n | r_{n-1})$ is the transition probability and $P(x_n | r_{n-1}, x_{n-r:n})$ is the data segment likelihood probability. We calculate the transition probability using equation

$$P(r_n | r_{n-1}) = \begin{cases} h(r_{n-1} + 1), & \text{if } r_n = 0 \\ 1 - h(r_{n-1} + 1), & \text{if } r_n = r_{n-1} + 1 \end{cases}$$

where $h(x) = g(x) / \sum_{i=x}^{\infty} g(i)$. We calculate the posterior probability using Bayes’s rule:

$$P(r_n | x_{1:n}) = \frac{P(r_n, x_{1:n})}{\sum_{i=0}^{n-1} P(r_i, x_{1:i})} \quad (4)$$

We calculate the posterior probability of the run length at that time index which corresponds to a new data sample. The pseudo code of this procedure is summarized in Algorithm 1.

4.3 Classification

We classify the sleep states using an online gradient descent method which leverages the importance weight on streaming data samples. To build up our classification model accurately we consider other sleep contexts such as body movements, but most of the sample data points resemble stationary states during the sleeping. Online gradient descent with importance weight aware updates [Karampatziakis and Langford 2011] helps to overcome this limitation of data by assigning weight

ALGORITHM 1: Change Point Detection

-
- 1: Initialize: $P(r_0) = 1$
 - 2: **for** Each new data point x_n **do**
 - 3: Calculate the data segment likelihood probability,
 $P(x_n|r_{n-1}, x_{n-r}:n)$
 - 4: Calculate the transition probability, $P(r_n|r_{n-1})$
 - 5: Calculate the joint distribution, $P(r_n, x_{1:n})$
 - 6: Find the posterior distribution on current run length, $P(r_n|x_{1:n})$
 - 7: Calculate the predictive distribution of x_n based on previous observation. $P(x_n|x_{n-1})$
 - 8: **end for**
-

to classes with lesser data points. The key principle here is: *The assignment of importance weight h to a sample that make it appears like a regular example of h times in the dataset.* We assume C is our classification model and use squared loss function for examining the consistency of C . The goal of our classification model is to minimize the loss function which reflects better accuracy. After each iteration of gradient descent, C is not altered, rather it is improved by adding an estimator h in order to optimize the loss function. We assume y is the true label and p is prediction of our model, where $l(p, y)$ is the loss function as shown in Eqn. (5). At each step C is updated using Eqn. (6)

$$l(p, y) = \frac{1}{2}(y - p)^2 \quad (5)$$

$$C_{m+1} = C_m + h(x) \quad (6)$$

Let w be the vector of weights and the training set is a set of (x_i, y_i, h_i) , $i = 1, \dots, T$ where x_t is a vector of d features. For linearity we assume $p = w^T x$. Our goal is to assign w in such a way so that the model C converges to the optimized solution. Assigning weight to a data point (x, y) , h times in a row have a cumulative effect with scaling factor $k(h)$ as shown in Eqn. (7). This scaling factor is defined by Eqn. (8) where η is the learning rate [Karampatziakis and Langford 2011]. At each iteration this weight is updated accordingly to the loss function l . Our base classifier C is a multinomial logistic regression model. Our proposed classification algorithm for finer non-stationary sleep states detection is shown in Algorithm 2. We first initialize the importance weights for each of the data instances and then train our classifier C . The weights of the data instances get adapted based on the prediction made by our trained model.

$$w_{i+1} = w_i - k(h)x \quad (7)$$

$$k(h) = \frac{p - y}{x^T x} (1 - e^{-h\eta x^T x}) \quad (8)$$

4.4 Active Learning Based Community Scaling

Our goal in this paper is to scale the sleep monitoring model to a community of individuals. While a significant research has been done on sleep monitoring and assessment and intervention strategies, lack of novel scaling algorithms prohibits the deployment, large scale validation, and acceptance of these technologies for healthy lifestyle, smart health and independent living applications. In this section we investigate how active learning based machine learning algorithms help build an informative model in presence of a minimal labeled datasets. We also depict how change point detection based time-series data analytics methodology help reduce the data uncertainty and guide to the selection of most informative query.

ALGORITHM 2: Importance Weighted Sleep Classification

```

1: Input: Extracted feature vectors from raw data with their respective labels.
2: Output: Trained model and updated importance weight of the data points.
3: Initialize:  $\forall_i w_i \leftarrow 0$ 
4: Get the feature vector for data point  $x_i$ 
5: while true do
6:   Train classifier  $C$ 
7:   Calculate the scaling factor  $k(h)$ 
8:   for  $i = 0$  to  $N$  do
9:     Calculate the weight  $w_i$  for each  $x_i$ 
10:    Update:  $w_i \leftarrow w_i - \frac{p_i - y_i}{x_i^T x_i} (1 - e^{-h\eta x_i^T x_i}) x_i$ 
11:   end for
12:   if  $l(p, y)$  converges then
13:     break
14:   end if
15: end while

```

Active learning has been proved to be very effective by combining it with supervised learning when a large pool of unlabeled data is available. Though traditional passive learning takes the initiative to label the unlabeled data randomly, but most of the data points which are selected randomly does not ensure better classification. It is difficult to collect all of the sleep related ground truth information from the user though by using the accelerometer sensor it is possible to broadly monitor the user sleep behavior and the specific sleep duration. To collect more fine-grained details about the sleep we train our proposed gradient based classifier with the causes of sleep disruption (such as waking up from nightmare, muscle cramp etc). By applying active learning, we propose to collect the labels of these informative data points so that our model can better classify the sleep stages and conditions and help scale this model in presence of minimal amount of ground truth. While applying active learning, one constraint is we have to assure that the whole labeling process doesn't become too intrusive. Crowd-sourcing can help us overcome this constraint by collecting a large amount of labeled data via arbitrary participants and provides aid in community scaling.

4.4.1 Query selection. In the following we briefly discuss the query selection approaches for active learning:

- **Query Synthesis:** The active learner asks the human annotator for "label membership" by using membership queries. In this approach the learner generates instances rather than sample from existing unlabeled set. But the problem with this approach is human annotator may have difficulty interpreting and labeling arbitrary instances.
- **Stream based selective sampling:** Each unlabeled instance is drawn at a time from the input source and the learner may decide instantly whether to query the instance or not. As we are using online classification algorithm and the data are processed in stream, we use this sampling strategy for our active learner.
- **Pool based sampling:** Evaluates and ranks the entire collection of unlabeled data before selecting the best query from a pool of instances.

4.4.2 Sampling metrics. Different sampling metrics such as least confident, margin sampling or maximum entropy based sampling are common in active learning algorithms. We propose to use the importance weighted active learning approach to build our community scaled sleep monitoring model [Beygelzimer et al. 2009]. To decide which points are most informative, we first calculate

the utility measurements of unlabeled data points. Whether a data point x_t will be queried or not depends on the history of labels seen so far based on our change point detection, gradient based classification and the identity of the point. If a change point is detected at data point x_t at time index t_n , and the label of x_t is inconsistent with the label of current run r_n , we invoke active learning. A probability measure p_t is maintained for each data point x_t . A coin flip, $Q_t \in \{0, 1\}$ with $E[Q_t] = p_t$ determines whether the data point will be queried or not. If the data point is queried based on the past history, then we update importance weight by $\frac{1}{p_t}$.

The active learning algorithm maintains an effective hypothesis space H_t through out the process. Initially, H_t contains all of the hypotheses from global space H . The expected loss of a hypothesis, $h \in H$ at time T is defined by Eqn. (9).

$$L_T(h) = \frac{1}{T} \sum_{t=1}^T \frac{Q_t}{p_t} l(h(x_t), y_t) \quad (9)$$

As it progresses, H_t becomes narrower by taking a subset, and ensuring that the factual loss of H_{t+1} is not much worse than the smallest loss, L_t^* in H_t .

$$H_{t+1} = \{h \in H_t : L_{t+1}(h) \leq L_t^*(h)\} \quad (10)$$

For each data point x_t , the active learning algorithm looks at the range of predictions and their losses by hypotheses in H_t and sets the sampling probability to the size of this range.

$$p_t = \max_{f, g \in H_t} \max_y l(f(x_t), y) - l(g(x_t), y) \quad (11)$$

If the range is too high than the rejection threshold then the hypotheses disagree greatly with each other. This certifies that the current prediction of x_t lies in the uncertain region. The active learning algorithm then queries for the label to settle the uncertainty. Our proposed active learning algorithm for largely reducing the micro-sleep states annotation effort is shown in Algorithm 3.

Apart from using only predefined class labels, the user can introduce new unseen class along with indicative attributes with the help of active learning. While prompting for label of data point x_t , we also collect the reason for their choice of label in restricted number of words. We find specific attributes from the provided reason and associate that attribute with the data point x_t . For example, if a user labels a data point as “getting up & sitting” and specifies the reason as “woke up from nightmare”, Sleep Well framework extracts the attribute “Nightmare” from the provided reason. Subsequently we re-evaluate our classification model and apply a recursive classification to associate the provided attributes to similar data points. This help our model to achieve microscopic sleep state classification, and finer evaluation for more elaborative and accurate diagnosis of patients and eventually scale the model beyond an individual premises.

4.5 Crowdsourcing

Crowdsourcing has been proved to be an effective component for collecting labels in many machine learning applications. Large scale data processing and annotating the data with multiple annotators or experts alleviate the traditional process for gathering ground truth data which is lengthy, costly, and time consuming. By using crowdsourcing though we accumulate a large volume of labeled data, but we also increase the risk of introducing a lot of noisy and ambiguous labels into our classifier. So it is necessary to identify potential reliable annotators and limit the effect of introducing noisy labels. On the other hand, a major challenge in crowdsourcing is to verify the provided labels. To tackle this we propose to calculate the inter-annotator agreement using Fleiss’ Kappa statistics and identify the proper label for the data point in concern. In our model we rank the annotators based on their reliability and awareness of the feedback. Here reliability refers to the correctness of

ALGORITHM 3: Active Learning with Importance Weighted Sampling

```

1: Input:  $L =$  set of labeled instances  $\{(x, y)^l\}_{l=1}^L$ 
    $U =$  set of unlabeled instances  $\{(x)^u\}_{u=1}^U$ 
   A classifier model,  $C_\theta$ 
2: Output: Updated classifier model,  $C_\theta$ .
3: Updated importance weight of queried data points.
4: for every instance in  $U$  do
5:   set  $p_t$  of instance  $x_t$  using equation (11)
6:    $y_t \leftarrow$  Prediction of  $C_\theta$  for  $x_t$ 
7:    $queried \leftarrow False$ 
8:   \* Check if  $x_t$  is a change point or not *
9:   if  $x_t$  falls in between successive run  $r_{n-1}$  and  $r_n$  using the posterior probability  $P(r_n|x_{1:n})$  then
10:    if  $y_t$  is not same as the label of current run  $r_n$  then
11:      query label  $y_t$ .
12:       $L_t \leftarrow L_{t-1} \cup \{x_t, y_t, \frac{1}{p_t}\}$ 
13:       $queried \leftarrow True$ 
14:    end if
15:  end if
16:
17:  if  $p_t$  is greater than rejection threshold and  $queried = False$  then
18:    query label  $y_t$ .
19:     $L_t \leftarrow L_{t-1} \cup \{x_t, y_t, \frac{1}{p_t}\}$ 
20:  else
21:     $L_t \leftarrow L_{t-1}$ 
22:  end if
23:  Update the hypothesis space  $H_t$ 
24: end for
25:  $C_\theta =$  Best hypothesis from  $H_t$ 
26: return  $C_\theta$ 

```

the feedback and awareness indicates the willingness of the annotators. For each annotator we maintain a probability measurement, δ_{tk}^j :

$$\delta_{tk}^j = P(y^j = k | y^j = t) \quad (12)$$

In equation 12, δ_{tk}^j denotes the probability that the annotator j provides the class label k to an instance given that the true class label is t . Our goal is to learn the parameter δ_{tk}^j for each annotator. Suppose there are R annotators. If the number of data instances is N , we initialize a $N \times R$ matrix M . M_{ij} denotes the label of instance i provided by annotator j . Given that y_i^t is the true label for instance x_i , we assume that $y_i^1, y_i^2, \dots, y_i^R$ are independent. Let $y = \{y_1^t, y_2^t, y_3^t, \dots, y_N^t\}$ be the set of true labels for the data instances. In our model δ is the reliability parameter. Our goal is to learn an optimal estimator \hat{y} of y to minimize the imposed error by provided noisy data points. By taking Beta prior on the reliabilities we can formulate a maximum likelihood estimator $\hat{\delta}$ of δ using equation 13. By taking y as hidden variable we can estimate $\hat{\delta}$ using expectation maximization [Raykar et al. 2010].

$$\hat{\delta} = \arg \max_{\delta} \log P(\delta | M) = \arg \max_{\delta} \log \sum_y P(\delta, y | M) \quad (13)$$

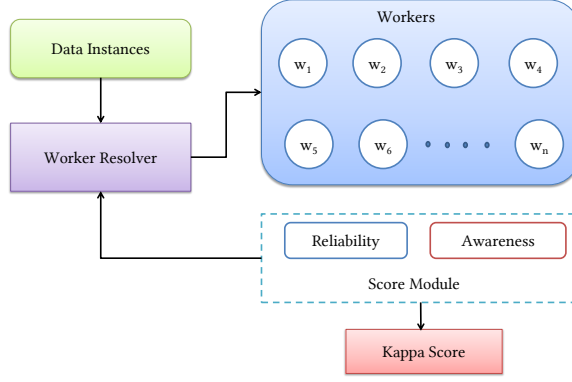


Fig. 2. An architectural overview of Crowdsourcing

We calculate the awareness of a certain class c , a_j^c by taking the percentage of data of class c labeled by j^{th} annotator. We assign weight w_j^c to each annotator by taking the product of their respective reliability of a certain data instance i and awareness. Based on the weight for each class we rank the annotators with respect to each class.

$$w_j^c = \delta_{tk}^j a_j^c \quad (14)$$

To verify our estimator we calculate the inter user agreement by using Fleiss' Kappa statistics. The kappa k is defined in equation 15. In equation 15, $1 - \bar{P}_e$ gives the degree of agreement that is achievable, and, $\bar{P} - \bar{P}_e$ gives the degree of agreement actually achieved. If the annotators agree with each other then $k = 1$, if not then $k \leq 0$.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (15)$$

Let n be the number of annotation per annotator, and let k be the number of classes. Then n_{ij} represent the number of annotators who assigned instance i to the j class. First calculate P_j , the proportion of all assignments which were to the class j :

$$P_j = \frac{1}{Rn} \sum_{i=1}^R n_{ij}, \quad 1 = \frac{1}{n} \sum_{j=1}^k n_{ij} \quad (16)$$

Now calculate P_i , which denotes how many annotator pairs are in agreement, relative to the number of all possible annotator pairs.

$$\begin{aligned} P_i &= \frac{1}{n(n-1)} \sum_{j=1}^k n_{ij}(n_{ij} - 1) \\ &= \frac{1}{n(n-1)} \sum_{j=1}^k (n_{ij}^2 - n_{ij}) \\ &= \frac{1}{n(n-1)} [(\sum_{j=1}^k n_{ij}^2) - (n)] \end{aligned} \quad (17)$$

Now we can measure \bar{P} , which is the mean of all P_i and \bar{P}_e using P_j :

$$\begin{aligned}\bar{P} &= \frac{1}{N} \sum_{i=1}^N P_i \\ &= \frac{1}{Rn(n-1)} \left(\sum_{i=1}^R \sum_{j=1}^k n_{ij}^2 - Rn \right)\end{aligned}\quad (18)$$

$$\bar{P}_e = \sum_{j=1}^k p_j^2 \quad (19)$$

In figure 2 the architecture of our crowdsourcing platform is shown.

5 SLEEP WELL FRAMEWORK EVALUATION

To evaluate our framework we focus on the following specificities. i) The performance of different classification algorithms in comparison to our classification approach, ii) Cross-user performance by building model with a user's sleep model and testing with someone else's model, iii) Performance of our framework using different wrist-band devices with accelerometer sensor, iv) Impact of active learning on our model, v) Precision of classifier when sleep attributes are introduced in the model by active learning.

5.1 System Implementation

We have implemented and tested our model by using two separate devices - wActiSleep-BT [Actigraph 2004] and EZ430-Chronos [EZ430-Chronos 2013], both of these devices contain 3-axis accelerometer sensor. EZ430-Chronos device also has heart monitor, pressure and temperature sensor. We collected raw accelerometer data from both of these devices using API provided by the manufacturers. We implemented our own software to extract raw data using C# programming language and then extracted the features using python numpy library. We sampled the data in 60Hz frequency. For importance weighted classification and active learning we used the machine learning tool- Vowpal Wabbit [Vowpal Wabbit 2016].

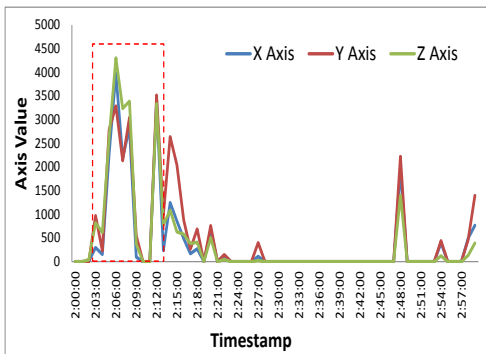


Fig. 3. Accelerometer reading when standing

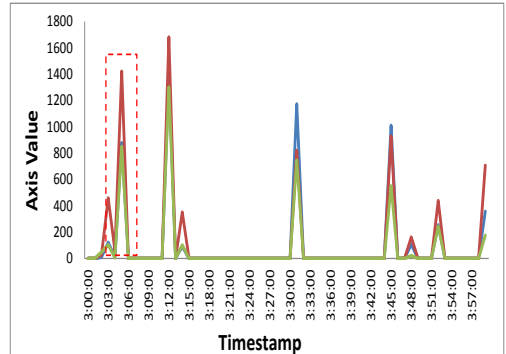


Fig. 4. Accelerometer reading when lying

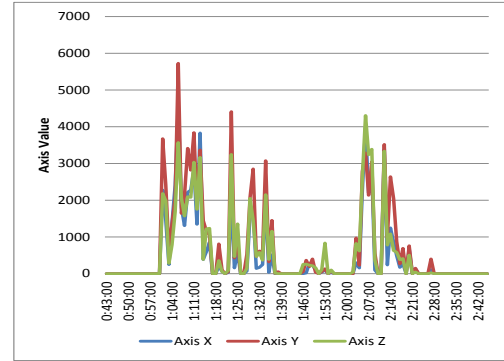
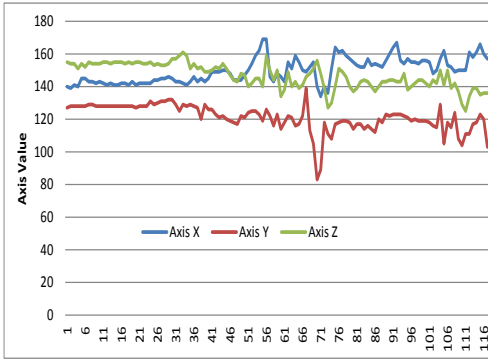


Fig. 5. Raw accelerometer data from Dataset 5.3.1. Fig. 6. Raw accelerometer data from dataset 5.3.2

5.2 Ground Truth Collection

We asked the users to log their sleep habits using sleep diary to correctly label the data points. We asked the participants to note down their sleep routines (preferred sleeping postures, regular hours of sleep, light intensity and sleep latency) each day of the experiment. There were many challenges involved while collecting the ground truth from the sleep diary. For example, consider two different scenarios, 1) the user is *awake & lying* and 2) *awake & not lying*. In case of stationary states (when the user is not moving but he is either lying or just sitting in the bed) the accelerometer readings are almost identical. Also when a user gets up in the middle of the night and performs some activities (checking his phone, going to bathroom etc.), there are movements involved. It was challenging to identify which movements were during sleep and which were due to some activities. The user was unable to correctly state the reason of movements in some cases. In Fig. 3 and 4, we can see two different movements (awake and standing, awake and lying). The user went to bathroom at 2:03 AM and came back to bed at 2:12 AM. On the other hand at 3:03 AM, the user was moving while lying. Therefore to assist the ground truth collection, we investigate a posture analysis using the inclination of the accelerometer. We observe that when perpendicular to gravity, accelerometers are more sensitive to small changes in inclination, but as the inclination increases the accelerometer becomes less sensitive to it. To resolve this issue we propose to use two axis. As we are using wrist worn bands, inclination of axis y and z are used to define the posture of the user. The z axis measures the direction of the gravity in the horizontal position, so coupling with the inclination of x axis help infer the posture of the user. We calculate the inclination of the device by using Eqns. (20) and (21).

We faced a challenge to define the threshold values for these inclinations as different users have different sleeping postures. We experimented with different sleep positions (On side, Face down, On your back) and calculated the inclination of the device in those positions. We ran the J48 decision tree classifier on the postural data. Based on the results of the classifier we defined the inclination threshold for different states, such as if $\theta_y < 16^\circ$ then the user is standing, if $16^\circ < \theta_y < 61^\circ$ the user is considered sitting, and for $\theta_x > 61^\circ$ the user is considered lying. Fig. 9 shows the results of our posture calculation using the inclination method. We also installed couple of motion sensors in the environment to strengthen our ground truth collection. We put two motion sensors (Aeotec Multisensor [Aeotec Multisensor 2006]) near both sides of the bed and another one mounted near the user's body when he/she is sleeping. The sensor mounted near the body captures the motion when the user is in the bed while the other two on the sides of the bed monitor when the user is out of the bed. While extracting information from the sensor, we assumed that consecutive two

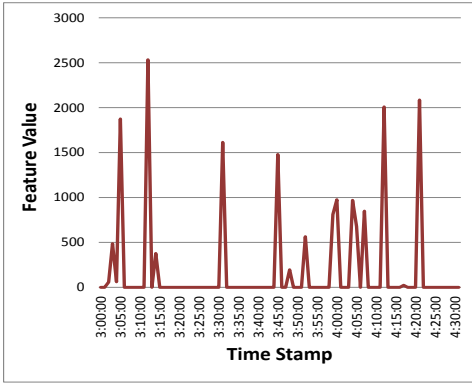


Fig. 7. Timestamped feature value.

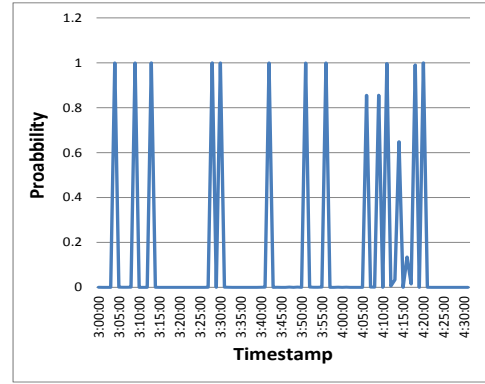


Fig. 8. Detected change points associated to figure 7

$$\theta_y = \tan^{-1}\left(\frac{y}{\sqrt{x^2 + z^2}}\right) \quad (20)$$

$$\theta_z = \tan^{-1}\left(\frac{z}{\sqrt{x^2 + y^2}}\right) \quad (21)$$

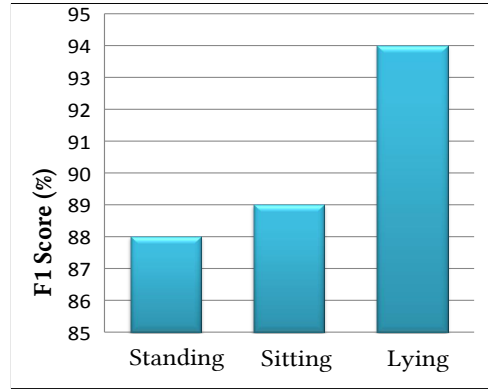


Fig. 9. Inclination measurement accuracy

data points from the sensors mounted on the sides correspond to getting in and then getting out of the bed. These multisensors also have built in light sensor, so we can detect the light condition in the sleeping environment using these sensors. Now we are able to validate the movements of the users and calculate the overall time he/she remained out of the bed efficiently by consolidating inclination measurement, motion sensor and data from sleep diary. We were able to label most of the data points correctly and remove noisy data points.

5.3 Datasets

We use real data traces collected from ≈ 60 users to validate the performance of our Sleep Well framework. We also compare our results for data from different body position.

5.3.1 Dataset with Clinical Ground Truth. We evaluate our model using publicly available benchmark dataset from Technische Universität Darmstadt [Borazio et al. 2014] which provides sleep phases determined by clinical polysomnography. The data set consists of timestamped raw acceleration data collected using wrist worn data logger at a sampling rate of 100Hz and includes the sleep stages (movements, awake, NoREM 1-3, REM, unknown) from 42 lab patients. The trend of raw accelerometer reading in this dataset is shown in Fig. 5. The sampling frequency is set to high due to the nature of the activities. Different sleep stages (NoREM 1-3 and REM) actually refer

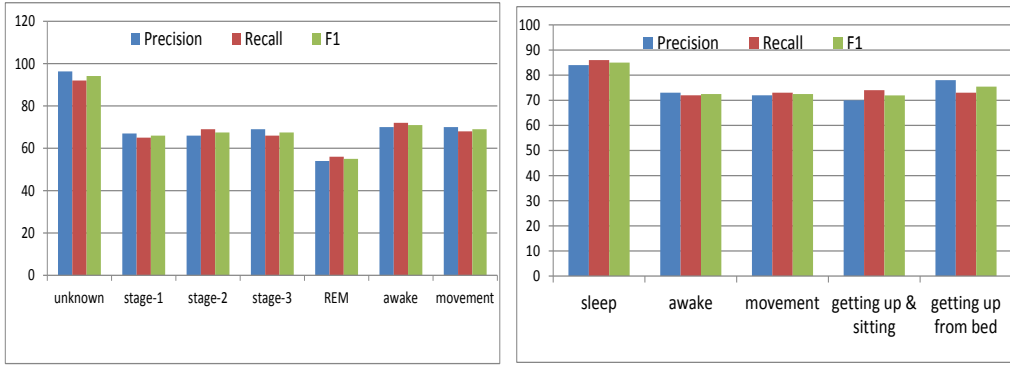


Fig. 10. Precision, recall and F1-measurement for Fig. 11. Precision, recall and F1-measurement with inter user classification (dataset 5.3.1) inter user classification (dataset ??)

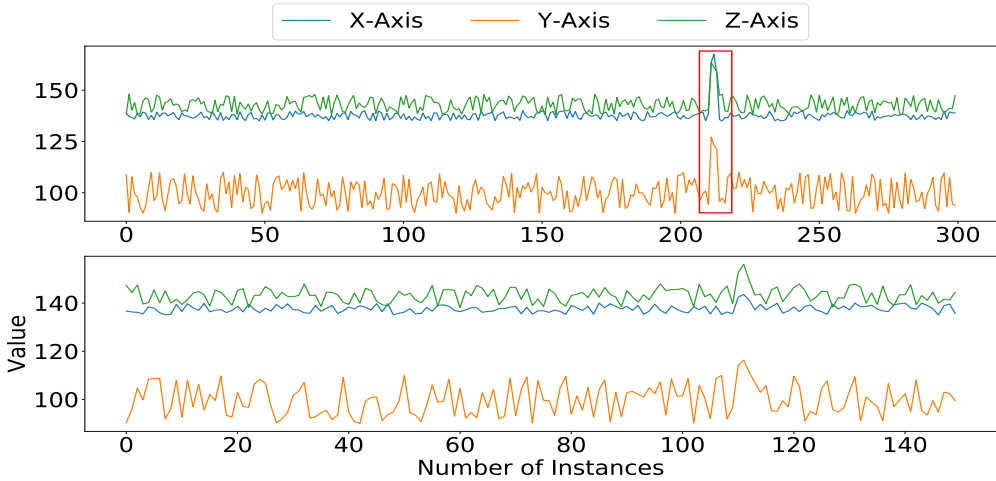


Fig. 12. The figure illustrates the effect of various sampling frequencies of the raw accelerometer data. The upper figure depicts raw accelerometer data at 100Hz frequency and the bottom figure at 25Hz frequency. The highlighted portion in the upper figure indicates a very subtle movement while sleeping.)

to macro activity - *sleep*. In order to understand the difference between movements pertaining to different sleep stages it is necessary to capture data at higher sampling frequency. We demonstrate the effect of high and low sampling frequency in Figure 12. The spike in the highlighted box in the upper figure indicates a very subtle movement while the user is sleeping. However, when downsampling the data to 25Hz (bottom figure) we experience no such spike and we lost the movement information. As a result it is necessary to choose a sampling frequency according to the definition and sensitivity of the classes. There are seven different classes in this dataset among which majority of the data points are labeled as unknown (51%) and awake (24%) with only a few important data points which affect the classification model. After inspecting the dataset, we note that the value of different data points of different classes were very close which imposes bias in our classification model. We handle this bias by assigning less weight to abundant data points (unknown and awake) and improve the classification process and accuracy.

5.3.2 Actigraph and Chronos Dataset. We collected sleep data using wActiSleep-BT and EZ430-Chronos at a sampling rate of 60Hz from 17 participants for two weeks. Out of 17 participants (11 males and 6 females) 13 were graduate students, 3 working professionals and 1 unemployed person. We conducted a survey beforehand to know about their sleep routine. Using the survey we gathered information regarding sleep time, average sleep hour, movement frequency in a scale of 1-5 and existing sleep disorders. We then selected a set of participants with diverse sleep routine and disorders. We asked the participants to put on the sensor when they go to the bed. The participants were also instructed to maintain a log the timing of getting up and getting in bed. The participants put the sensors on their waist using a belt. We noticed that wActiSleep-BT device has better sensitivity due to slight movements rather than EZ430-Chronos which help differentiate between actual movements and sleep patterns from a user. Almost 65% data points of this dataset belong to *Sleep* class and 22% to *Awake* class. As a consequence, our dataset is also imbalanced. Fig. 6 shows the raw readings from ActiSleep device.

5.4 Evaluation Methodology

5.4.1 Supervised Learning. We carried out our experiments with 17 participants (11 males and 6 females) over two weeks where each participant has provided data for 8-10 days. We validated that 17 is a statistically significant population size using *t-test*. The trend in training accuracy (intra user) with respect to size of the population is shown in figure 13. We see that after increasing the population size more than 17, the accuracy is not changing significantly. We proved our hypothesis that it did not happen by a chance by conducting *t-test* using dataset 5.3.1. We conducted *t-test* with varying population size and received *p*-value of 2.021 with 95% confidence interval. Although we get a bit more accuracy if we increase the population size (74%), however due to training time and resource consumption we chose 17 as the optimum population size. Out of these 17 participants, 7 wore the EZ430-Chronos device and other 10 put on wActiSleep-BT. We split each data set into two parts, one for training and other for testing. To overcome the class imbalance problem, the importance weight play a significant role. We look at the confusion matrices of classifying two classes (Sleep, Awake) in Table 2 & 3. It is evident that due to class imbalance, a lot of the Awake class instances are inferred as Sleep if importance weight is not used. We applied our classification models on the datasets mentioned in sections 5.3.1 and 5.3.2.

	Sleep	Awake
Sleep	91%	9%
Awake	14%	86%

Table 2. Confusion matrix of Sleep/Awake classifier with importance weight.

	Sleep	Awake
Sleep	80%	20%
Awake	51%	49%

Table 3. Confusion matrix of Sleep/Awake classifier without importance weight.

Intra User Classification. We tested different classification models with our proposed Online Stochastic Gradient Descent (OSGD) method - Support Vector Machine (SVM), Multilayer Perceptron (MP), LogitBoost (LB), Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT) - using different user's dataset. The accuracy of different classification model using one of the subject's dataset from each datasets is shown in Table 4 and 5. The average accuracy of OSGD is 73.20% for a patient from dataset 5.3.1 and 78.45% for dataset 5.3.2. This attests that consideration of inclination and sensor data and using it to correct labels in dataset 5.3.2 help yield better classification results. Also the results indicate that putting the device on the waist endows better accuracy. We investigated this disparity, and found that hand movements are more abrupt and arbitrary which

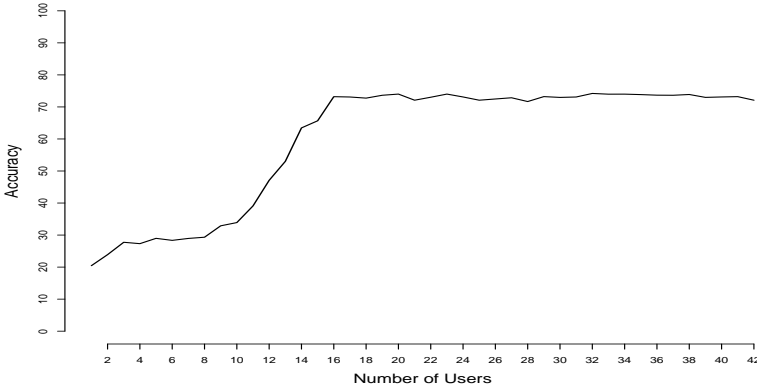


Fig. 13. The trend in intra user classification accuracy with varying population size.

Table 4. Accuracy (Dataset : 5.3.1)(%) (Intra User)

	OSGD	SVM	MP	DT	LR	LB	RF
Unknown	98.76	96.50	85.80	99.01	98.93	98.37	97.95
Stage-1	69.45	44.44	60.36	58.57	47.63	61.82	70.40
Stage-2	70.29	41.02	58.54	59.09	48.18	68.46	71.87
Stage-3	68.36	39.15	63.36	48.24	49.33	60.77	63.94
REM	58.22	37.28	49.11	41.55	38.31	41.03	59.10
Awake	74.78	68.12	72.10	70.01	64.96	66.90	72.73
Movement	72.59	69.31	62.88	70.66	65.60	63.27	71.25
Average	73.20	56.54	66.13	63.87	58.99	65.80	72.46

Table 5. Accuracy (Dataset : 5.3.2)(%) (Intra User)

	OSGD	SVM	MP	DT	LR	LB	RF
Sleep	87.79	87.98	80.8	84.01	73.32	76.63	88.52
Awake	77.9	71.35	75.66	67.91	73.56	70.21	75.69
Movement	76.25	74.87	68.32	68.41	70.27	72.11	75.14
Getting up & sitting	72.11	64.58	67.39	68.11	64.85	69.15	70.02
Getting up from bed	78.21	69.89	70.36	70.1	71.19	62.39	73.39
Average	78.45	73.73	72.50	71.70	70.63	70.09	76.54

results in more confusing data points. Also very subtle body movements are difficult to distinguish when using a wrist worn accelerometer.

The major accuracy improvement was noticed for inferring the micro sleep state. Although individual accuracy for classes - *Stage 1*, *Stage 2* and *REM* for Decision Tree (DT) classifier was better in dataset 5.3.1, but the average accuracy of inferring sleep states (sleep stage 1-3, REM) is 66.58% which is better than the average of DT classifier (66.32%). While for our dataset we achieved 87.79% accuracy.

Cross User Classification. It is important that a classification process will not only recognize the sleep states of an already seen user, but also help generalize the classification for new users. We cross validated our approach with inter user classification model. We trained our model using 20 patient’s data from dataset 5.3.1 and tested the trained model with remaining 22 patient’s data. The average accuracy was 69.79%. With data from dataset 5.3.2 we achieved 75.46% overall accuracy. Fig. 10 and 11 shows the results in Precision (percentage of times that a recognition result made by the model is correct), Recall (percentage of times that a sleep state is detected) and F1-measurement (combination of both recall and precision) for both the datasets. Also in Fig. 14 represents the trend of loss for different datasets.

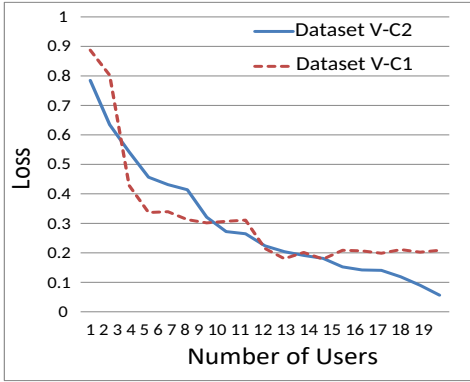


Fig. 14. Trend of loss for inter user classification in different datasets

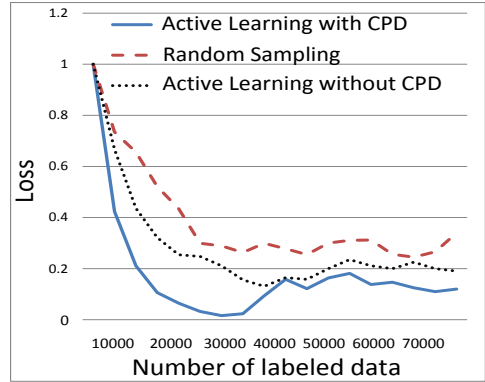


Fig. 15. Different active learning techniques for Dataset 5.3.1.

5.4.2 Active Learning Experiments. In addition to supervised learning, we evaluate how we can improve the classification result using active learning with minimal user feedback. We have discussed our active learning algorithm in section 4.4. We sampled both the datasets with a window of 60 seconds on accelerometer data. Each sample is a feature vector with 16 dimensions. Initial labeled dataset L_1 consisting of 135089 samples (from dataset 5.3.1) and L_2 consisting of 42,000 samples (from dataset 5.3.2) are provided to the individual classifier C_1 and C_2 for training. Then unlabeled dataset U_1 of 510113 (dataset 5.3.1) and U_2 of 121,147 samples (dataset 5.3.2) are used to test the classifier C_1 and C_2 . The samples are provided sequentially with respect to timestamp.

The uncertain data points, meaning the points which the classifier was unable to classify are queried in accordance with our active learning algorithm (3). We calculated the loss at each phase after a data point is queried and the model is re-trained. We compared our result with randomly selected samples for labeling. To further assist the active learning process we validated the results with our change point detection (CPD algorithm discussed in 4.2). When a change point is detected in the dataset, we cross validated the change points with the classification result with respect to the timestamp. Figs. 7 and 8 plot the association of change points with timestamped accelerometer data points. If the label of the sample is not consistent between each of the model we imposed active learning and queried the data point. Initially with L_1 and L_2 we note the average classification accuracy as 63.8% and 70%. We applied importance weighted active learning, and see that the model converged faster with change point detection. 86719 samples (17% of total samples) from U_1 and 8843 samples (7.3% of total samples) from U_2 were queried for the model to converge in presence of CPD which helped achieve 72% (dataset 5.3.1) and 76.89% (dataset 5.3.2) accuracy, while with

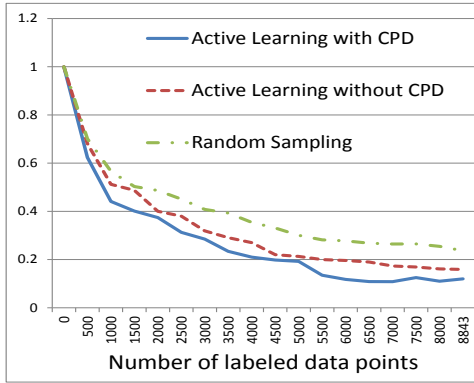


Fig. 16. Different active learning techniques for Dataset 5.3.2.

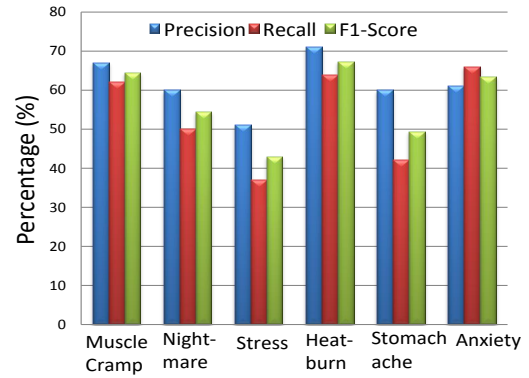
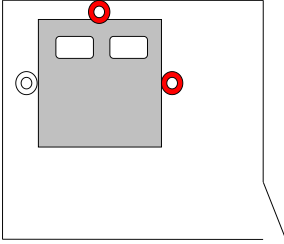


Fig. 17. Different attributes classification result in getting up & sitting partition.

randomly selected data points 68% and 73% accuracy was observed. Fig. 15 and 16 shows the change in loss with random sampling, active learning with and without CPD techniques with different datasets. We see that active learning with CPD outperforms the other strategies. In case of dataset 5.3.1 we notice from Fig. 15 that the change in loss is irregular. After analyzing the dataset 5.3.1 we found out that due to the presence of noisy data points the loss increased.

5.4.3 Crowdsourcing Experiments. During this process we faced a challenge regarding what kinds of data to show which can reflect the sleep classes. As audio, video or image data violate the privacy of the user so we had to come up with a different methodology rather than traditional image based crowd-sourcing. We presented some semantic information from the users sleeping habit (regular hours of sleep, sleep latency, posture, average number of times the user gets up at night, how much the user moves on average in percentage and light condition) and a visual illustration of sensor activation (discussed in section 5.2) to the annotators. In Fig. 18, we show an example of visual illustration. The double circled objects represent sensors and the activation is marked by red color. In this example the sensor mounted near the head and the sensor mounted near the right side of the bed are activated as the user was getting up from the bed. 10 annotators participated in our crowdsourcing experiment with a dataset containing 10,000 data instances from Dataset 5.3.2. In table 6 the kappa coefficient, z-score and p-value for individual classes are shown. The kappa coefficients of sleep, awake and movement classes are considerably low than other two classes. This was due to the nature of these activities and sensitivity of the motion sensor. During our experiment we have seen that, movements with smaller intensities while in sleep or awake is sometimes not captured by the motion sensor. As a result most of the annotators, defined those data instances as sleep. On the other hand, movement with higher intensities while sleeping are annotated as awake, movement and getting up & sitting. Also the first three activities are related to the state lying. As a result the posture information using inclination measurements did not help much. For getting up from bed the visual illustrations

were less noisy and easier to depict, as a result the inter user agreement score is much better.



Class	Kappa	Z-Score	P-Value
Sleep	0.468	2.8	0.167
Awake	0.423	5.97	0.013
Movement	0.447	4.52	0.00678
Getting up & Sitting	0.537	2.298	0.051
Getting up from the bed	0.835	1.256	0.0006

Table 6. Fleiss Kappa Score (Inter user agreement).

Fig. 18. Visual illustration of sensor activation.

5.4.4 Introduction of New Unseen Class and Attributes. A user is able to personalize the model by introducing new unseen classes and attributes with the help of active learning. We simulated our active learning algorithm by introducing new class labels in the classification model. While collecting the query label we also asked for the reason behind choosing the label from the annotator, so that we can look for important indicators for the clinicians. We restricted the length of the reason in 5 words. For example, if a sample is queried and the annotator labels the sample as *getting up and sitting*, he can also state the reason for labeling the data such as *muscle cramp, stress or anxiety, nightmare* etc., which are microscopic events for sleep disruption. We applied a nested classification by considering these microscopic events as class labels. After classifying using our defined general class labels, we partitioned each class label data and applied our classification algorithm in separate partitions again by considering the provided attributes as labels. For example, let us assume an user states reason ‘A’ as the cause of sleep disruption or any kinds of changes in the pattern. Our framework then partitions the data and the number of partition is equal to the number of class labels (in our model it is 5), as a result in each partition the data points are of same class. Sleep Well framework then performs a classification on separate partitions with class label ‘A’. This nested classification process ascertain the microscopic sleep events. The precision, recall and F1 score of recorded attributes (muscle cramp, heatburn, stomach ache, stress, anxiety, and nightmare) for parent class “*getting up and sitting*” is presented in Fig. 17.

6 SLEEP SCORING

Various sleep scoring models have been proposed over the years like PSQI [Buysse et al. 1989], Webster [Webster et al. 1982], Sleep Quality Scale [YI et al. 2006] Cole et al. [Cole et al. 1992], Sadeh [Sadeh et al. 1994], ‘ZQ’ of Zeo [Zeo 2003] which takes in consideration total sleep, deep sleep, REM sleep, wake time and number of times woken up. Webster [Webster et al. 1982], Cole-Kripke [Cole et al. 1992], Sadeh [Sadeh et al. 1994], Jean Louis et al [Jean-Louis et al. 1997] models use the knowledge of actigraphy. In actigraphy the presence of movements indicates wakefulness and the absence of movements indicates sleep. Sleep efficiency is then calculated by taking the ratio of time slept versus total time spent in bed. PSQI [Buysse et al. 1989] is designed for assessing long term sleep quality. PSQI contains nineteen questions regarding the habit of sleep and the events which cause sleep disruption. The habit of sleep includes sleeping time, sleep latency, sleep duration and wake up time. From our classification results we can easily find out the trend of user’s sleep (sleeping time, wake up time, sleep duration). In our framework we are collecting the reasons for sleep disruption using active learning which also help in answering the questions related to trouble in sleep. In our work, we calculate the sleep wake cycle of the user using Cole-Kripke [Cole et al. 1992] algorithm to verify the sleep duration and try to find answers to as many questions as possible from PSQI [Buysse et al. 1989].

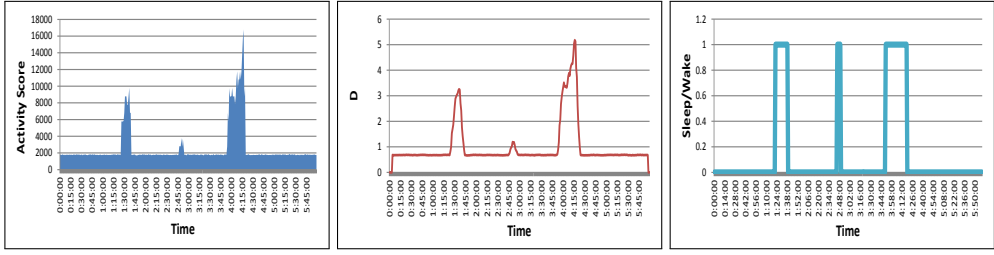


Fig. 19. Activity score for times- Fig. 20. Calculated D from corre- Fig. 21. Sleep/Wake estimation tamps between 12:00 AM - 6:00 AM
 AM 23 AM - 6:00 AM

In Cole-Kripke algorithm the sleep/wake state at any epoch is calculated by considering previous four minutes and next two minutes actigraphy information. The model is defined by the following equation:

$$D = P(W_{-4}A_{-4} + W_{-3}A_{-3} + W_{-2}A_{-2} + W_{-1}A_{-1} + W_0A_0 + W_{+1}A_{+1} + W_{+2}A_{+2}) \quad (22)$$

Here P denotes the scaling factor, W_{-i} , W_0 , and W_{+i} are the weighting factors for previous minute, present minute and the following minute. The activity scores for the previous, present and following minute are expressed by A_{-i} , A_0 , and A_{+i} . If $D < 1$ then the state is considered to be sleep and for $D \geq 1$, wake. We will adapt the solution for these parameters proposed in [Cole et al. 1992] developed by [Webster et al. 1982]. The rule of assigning scores are - (a) after 4 minutes scored as wake, next 1 minute is also scored as wake (b) after 10 minutes scored as wake, next 3 minutes are also scored as wake (c) after 15 minutes scored as wake, next 4 minutes are scored as wake (d) 6 minutes or less surrounded by at least 10 minutes scored as wake (before and after) are also scored as wake and (e) 10 minutes or less surrounded by at least 20 minutes scored as wake (before and after) are also scored wake. Then regression was applied for solving 22 using the scores calculated from dataset 5.3.1. After fitting the values in Eqn 22, we get the following equation:

$$D = 0.00001(404A_{-4} + 598A_{-3} + 326A_{-2} + 441A_{-1} + 1408A_0 + 508A_{+1} + 350A_{+2}) \quad (23)$$

Where activity score A denotes the number of epochs with movements in that particular minute. We calculate the state of each minute using 23 and find the number of minutes the user was sleeping. After we calculate the following components of PSQI:

- (1) We identify the usual time to go to bed at night from the user log diary and validating it with the mounted motion sensor firing sequence. Our framework maintains two separate variables for the motion sensors mounted on the sides of the bed - r and l . Upon firing of the sensors we increase the variables. If $\frac{|r-l|}{2}$ is even then the user got in and then got out of the bed and if odd the user is in in the bed.
- (2) Usual slack time for the user to fall asleep by subtracting the timestamp of answer (a) from the usual time the user fell asleep. Additionally we record the slack time for each day.
- (3) Usual time the user gets up in the morning.
- (4) Hours of sleep at each night.
- (5) We calculate the number of days the user could not sleep within 30 minutes using the information in (b).
- (6) How many times the user woke up in the middle of the night or early morning using our classification model and the motion sensor data.

- (7) How many times the user got up to use the bathroom using active learning and attribute detection (section 5.4.4).
- (8) Cannot breathe comfortably using active learning.
- (9) feel too cold and too hot using active learning.
- (10) Have bad dreams or nightmare using active learning.
- (11) Have pain using active learning.
- (12) Other reasons for the sleep disruption using active learning.

Rest of the components of PSQI are subjective which we could not measure using our framework. But we are able to automate most of the components with proper validation. Queries 8-12 are not posed in real-time. We pose these queries later in the following day. In figure 19, 20 and 21, the activity score, D score and sleep/wake state estimation from D score has been plotted. From the figures, it is visible that the participant was awake thrice. We verified our findings using participant's log diary where it was stated that at around 1:30 AM and 3:45 AM he woke up for using the restroom. Around 2:45 AM we see a spike in and the motion sensor near the head also fired but the participant did not mention anything about staying awake during this time. After calculating the PSQI components we calculate the final PSQI score. Figure 22 and 23 shows the frequency distribution of PSQI scores using actigraphy and the original questionnaire for 17 participants respectively. Although the distributions look quite similar, but the population size of 17 is not statistically significant. We want to investigate and prove our hypothesis with a larger population size in future.

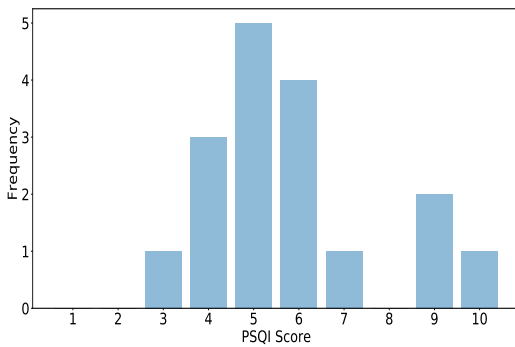


Fig. 22. Histogram of PSQI scores calculated using actigraphy.

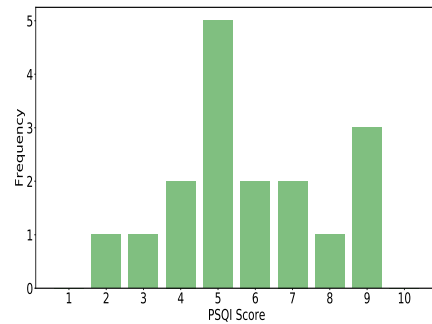


Fig. 23. Histogram of PSQI scores using PSQI questionnaire.

7 DISCUSSIONS AND FUTURE DIRECTIONS

In the current version of Sleep Well framework, we did not discuss about individual sleep scoring based on our sleep state classification. Most of the sleep scoring models like Pittsburgh Sleep Quality Index (PSQI) [Buysse et al. 1989] or Webster et al. [Webster et al. 1982] do not consider the habit of individual's sleep. In our experiment, we found that one participant was moving frequently while sleeping and woke up 2 or 3 times at night to use the bathroom. Even if the participant had disrupted sleep according to our data, but in accordance to participant's feedback he had a sound sleep. We look forward to investigate the sleep habits of individuals (like movements and getting up frequently) using change point detection in future and devise a dynamic sleep scoring module. Different components of PSQI such as PSQIDURAT (duration of sleep), PSQIDISTB

(sleep disturbance), PSQILATEN (sleep latency), PSQIDAYDYS (day dysfunction due to sleepiness), PSQIHSE (sleep efficiency), PSQISLPQUAL (Overall sleep quality), PSQIMEDS (Need meds to sleep) can be presented in the form of a tree data structure where each component represents a node in the tree. The outcome of the diagnostic procedures obtained from the set of questions being asked can help calculate the average score of each of the component and can be assigned as a conditional statement to each of the branch.

In current version of our framework, we invoke the change point detection only in case of active learning. As a future research direction, it is possible to combine change point detection with classification model to perform a more thorough time series analysis. For example, abrupt sleep disturbances (muscle cramp, nocturnal panic attacks etc.) cause sudden changes in the data points. It would be beneficial to capture these subtle changes and correlate the run length with the classification process.

In the current implementation, we considered only the inclination of the device to infer the user's current posture. Therefore, another future direction is to integrate the locomotive activity (such as sitting, standing, walking etc.) recognition with our framework for improving the noise reduction methodology. The orientation and the positioning of the devices affects the performance of the model as well. We would like to investigate transfer learning algorithm to address this issue in our future work. For active learning experiments we assumed that the user will always provide the correct label. In real life, it is possible that the user may provide wrong label or leave it blank. Also while collecting the attributes of sleep disruption causes, we are using a hard coded attribute list. If the labeler provides different attributes then the model fails to assimilate them. In such cases, imperfect annotation handling and optimal querying can be further studied to improve the performance of active learning. This is also applicable for crowd-sourcing the sleep data for malicious user identification. The visual illustration provided while applying crowd-sourcing was also not able to provide accurate representation in some cases. Although we provided the semantic information about the users sleep routine, but it did not improve the annotation performance as most of the annotators ignored the information and provided feedback based on the visual illustration only. Without the ground truth information, this runs the risk of introducing noisy instances in the model. We want to investigate whether adding other modalities like highly sensitive motion sensor and smart home data (appliance usages) can improve the performance of the annotators in future. In our current state, we collect the reason of the label from the user and extract the important microscopic attributes from the provided reason. For future, these feedbacks can be further leveraged to inspect the nature of various sleep disturbances at the microscopic level which will greatly help in longitudinal sleep assessment and diagnoses.

7.1 Reliability of Sleep Technologies

The growing pervasiveness of Off-the-shelf sensor rich wearable and mobile devices in our daily lives presents the convenience to capture the underlying contextual information of activities of daily living inconspicuously. However the variations of these commercial devices with respect to device manufacturers, CPU powers and OS types pose serious challenges in the applicability of these technologies in the same settings across different domains. The major type of heterogeneity which is evident is the different sampling frequency of these devices [Stisen et al. 2015]. Also most of the off-the-shelf devices are designed to be appropriate for clinical studies. In our experiments the two devices that we used (Actigraph and Chronos) are research devices which provide much higher sampling frequency than the commercial wearable devices. Although the use of these devices in clinical studies is not reliable but these modalities can compliment the clinical study by providing extra evidences. For example we have shown in Section 6 that we can provide further evidences for some of the questions that the participants are supposed to answer for PSQI test [Buysse et al.

1989]. On the other hand we cannot always answer the subjective aspect of clinical studies using actigraphy as discussed in Section 6. The authors of [Lockley et al. 1999] also showed that by comparing the actigraphic and subjective measurement of sleep. They found that actigraphic sleep monitoring is inferior in calculating sleep/wake cycle, sleep latency, duration of night awakenings and sleep onset. Actigraphic approach misclassifies inactivity as sleep for example, just lying down and watching television and not moving much will be regarded as sleep and creates ambiguity in the overall calculation of sleep performance. From this perspective actigraphic approaches might seem unreliable, however we can overcome this problem by monitoring the daily sleep routine of a person and augment that information in our classification model. Based on these evidences we think that wearable devices are still not enough to replace the clinical sleep monitoring method, however they can certainly endorse other important evidences which can help to aid the clinical trials.

7.2 Reliability of Sleep Apps

Apart from wearable devices, there are couple of mobile applications which provide insights about sleep routines of the users. Sleep as Android [Sleep As Android 2010], Sleep Time Smart Alarm Clock [Sleep Time Smart Alarm Clock 2015], [Sleep Bot 2017] app provides information regarding sleep/wake cycle and sometimes about sleep environment as well. These statistics only provide rough estimates and ultimately experience lack of reliability for measuring the sleep performances. Recently these apps are also capable of extracting more evidences from paired wearables. Although this improves the inference capability, but does not ensure the improvement in reliability.

8 CONCLUSIONS

In this paper, we described the design, implementation and evaluation of Sleep Well, a sleep monitoring framework which helps classify the microscopic sleep states using wearable devices. We postulated a gradient descent-based approach which incorporates with importance weight aware updates in the microscopic sleep state detection process. We also consolidated our framework by blending change point detection and active learning in the inference pipeline. Our classification achieved 78% accuracy with the aforementioned experimental setup. The empirical results demonstrate the effectiveness of our framework in determining different sleep states. The result increased by 7% when active learning was employed. Our approach helps accelerate the faster convergence to optimal sleep states detection accuracy using minimal user feedback in presence of active learning. Besides that with the help of change point detection, we were able to validate and interpret the transitions between these sleep states. In future, we plan to investigate the combination of change point detection and classification to further improve the accuracy. Also by conforming the attributes from user provided feedback into our architecture will help provide meaningful insights for better understanding of sleeping behavior.

ACKNOWLEDGEMENT

This research was supported in part by the NSF under Grants CNS-1344990, CNS-1544687, and IIP-1559752; the ONR under Grant N00014-15-1-2229; Constellation: Energy to Educate; the UMB-UMBC Research and Innovation Partnership Grant and the Alzheimer's Association Research Grant AARG-17-533039.

REFERENCES

- Actigraph. 2004. <http://www.actigraphcorp.com/>. (2004).
- Ryan Prescott Adams and David J.C. MacKay. 2007. Bayesian Online Changepoint Detection. Cambridge, UK.
- Aeotec Multisensor. 2006. <http://aeotec.com/z-wave-sensor>. (2006).

- Hande Özgür Alemdar, Tim van Kasteren, and Cem Ersoy. 2011. Using Active Learning to Allow Activity Recognition on a Large Scale. In *Ambient Intelligence - Second International Joint Conference on Aml 2011, Amsterdam, The Netherlands, November 16-18, 2011. Proceedings*. 105–114.
- Android Wear. 2014. http://www.android.com/intl/en_us/wear/. (2014).
- Salikh Bagaveyev and Diane J. Cook. 2014. Designing and evaluating active learning methods for activity recognition. In *The 2014 ACM Conference on Ubiquitous Computing, UbiComp '14 Adjunct*. 469–478.
- Yin Bai, Bin Xu, Yuanchao Ma, Guodong Sun, and Yu Zhao. 2012. Will You Have a Good Sleep Tonight?: Sleep Quality Prediction with Mobile Phone. In *Proceedings of the 7th International Conference on Body Area Networks*.
- Basis Band B1. 2014. <http://www.mybasis.com/>. (2014).
- Beddit. 2015. <http://www.beddit.com/features/>. (2015).
- Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. 2009. Importance Weighted Active Learning. In *ICML*. 49–56.
- Marko Borazio, Eugen Berlin, Nagihan Küçükıldiz, Philipp M. Scholl, and Kristof Van Laerhoven. 2014. Towards Benchmark Sleep Detection with Inertial Wrist-worn Sensing Units. In *ICHI 2014*. Verona, Italy.
- D. Buysse, C. Reynolds, T. Monk, S. Berman, and D. Kupfer. 1989. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. In *Psychiatry Research*, Vol. 28. 193–213.
- Yung-Ju Chang, Gaurav Paruthi, Hsin-Ying Wu, Hsin-Yu Lin, and Mark W. Newman. 2017. An investigation of using mobile and situated crowdsourcing to collect annotated travel activity data in real-world settings. *Int. J. Hum.-Comput. Stud.* 102 (2017), 81–102.
- Zhenyu Chen, Mu Lin, Fanglin Chen, N.D. Lane, and et al. 2013. Unobtrusive sleep monitoring using smartphones. In *PervasiveHealth*. 145–152.
- Anand Inasu Chittilappilly, Lei Chen, and Sihem Amer-Yahia. 2016. A Survey of General-Purpose Crowdsourcing Techniques. *IEEE Trans. Knowl. Data Eng.* 28, 9 (2016), 2246–2266.
- Klerman EB Barbieri R. Citi L, Bianchi MT. 2011. Instantaneous monitoring of sleep fragmentation by point process heart rate variability and respiratory dynamics. In *Engineering in Medicine and Biology Society, 2011*.
- R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, and J. C. Gillin. 1992. Automatic sleep/wake identification from wrist activity. *Sleep* 15, 5 (Oct 1992), 461–469.
- Florian Daiber and Felix Kosmalla. 2017. Tutorial on wearable computing in sports. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2017, Vienna, Austria, September 4-7, 2017*. 65:1–65:4.
- Andrew L. Chesson et al. 1997. Practice Parameters for the Indications for Polysomnography and Related Procedures. *Sleep* 20 (1997), 406–422.
- EZ430-Chronos. 2013. <http://www.ti.com/tool/ez430-chronos>. (2013).
- Muhammad Fahim, LeBa Vui, Iram Fatima, Sungyoung Lee, and Yongik Yoon. 2013. A Sleep Monitoring Application for u-lifecare Using Accelerometer Sensor of Smartphone. In *Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction*. Vol. 8276. 151–158.
- Fitbit. 2007. <http://www.fitbit.com/>. (2007).
- N. Foubert, A.M. McKee, R.A. Goubran, and F. Knoefel. 2012. Lying and sitting posture recognition and transition detection using a pressure sensor array. In *Medical Measurements and Applications Proceedings (MeMeA)*. 1–6.
- Weixi Gu, Zheng Yang, Longfei Shangguan, Wei Sun, Kun Jin, and Yunhao Liu. 2014. Intelligent Sleep Stage Mining Service with Smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. 649–660.
- Tian Hao, Guoliang Xing, and Gang Zhou. 2013. iSleep: Unobtrusive Sleep Quality Monitoring Using Smartphones. In *SenSys*. Article 4, 4:1–4:14 pages.
- Hello Sense. 2014. <https://hello.is/>. (2014).
- Enamul Hoque and John A. Stankovic. 2010. Monitoring quantity and quality of sleeping using WISPs. In *IPSN*. 370–371.
- Enamul Hoque and John A. Stankovic. 2012. AALO: Activity recognition in smart homes using Active Learning in the presence of Overlapped activities. In *6th International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth 2012 and Workshops, San Diego, CA, USA, May 21-24, 2012*. 139–146.
- H. M. Sajjad Hossain, Md Abdullah Al Hafiz Khan, and Nirmalya Roy. 2017. Active learning enabled activity recognition. *Pervasive and Mobile Computing* 38 (2017), 312–330.
- H. M. Sajjad Hossain, Nirmalya Roy, and Md Abdullah Al Hafiz Khan. 2015. Sleep Well: A Sound Sleep Monitoring Framework for Community Scaling. In *16th IEEE International Conference on Mobile Data Management, MDM 2015, Pittsburgh, PA, USA, June 15-18, 2015 - Volume 1*. 44–53.
- Kyuwoong Hwang and Soo-Young Lee. 2012. Environmental audio scene and activity recognition through mobile-based crowdsourcing. *IEEE Trans. Consumer Electronics* 58, 2 (2012), 700–705.
- G. Jean-Louis, H. von Gizycki, F. Zizi, A. Spielman, P. Hauri, and H. Taub. 1997. The actigraph data analysis software: I. A novel approach to scoring and interpreting sleep-wake activity. *Percept Mot Skills* 85, 1 (Aug 1997), 207–216.

- George H. John, Ron Kohavi, and Karl Pflieger. 1994. Irrelevant Features and the Subset Selection Problem. In *Machine Learning, Proceedings of the Eleventh International Conference, Rutgers University, New Brunswick, NJ, USA, July 10-13, 1994*. 121–129.
- Murray W. Johns. 2000. Sensitivity and specificity of the multiple sleep latency test (MSLT), the maintenance of wakefulness test and the Epworth sleepiness scale: Failure of the MSLT as a gold standard. *Journal of Sleep Research* 9, 1 (2000), 5–11.
- Stephan Jonas, Andreas Hannig, Cord Spreckelsen, and Thomas M. Deserno. 2014. Wearable technology as a booster of clinical care. *Proc. SPIE* 9039.
- Darkhovsky B Fell J Kaplan A, Röschke J. 2001. Macrostructural EEG characterization based on nonparametric change point segmentation: application to sleep analysis.. In *Journal of Neuroscience Methods*. 81–90.
- Nikos Karampatziakis and John Langford. 2011. Online Importance Weight Aware Updates. In *Proceedings of the Twenty-Seventh Conference on Uncertainty (UAI) in Artificial Intelligence*. 392–399.
- Matthew Kay, Eun K. Choe, and et al. 2012. Lullaby: A Capture & Access System for Understanding the Sleep Environment. In *UbiComp*.
- Christopher Kline. 2013. *Sleep Quality*. Springer New York, New York, NY, 1811–1813. https://doi.org/10.1007/978-1-4419-1005-9_849
- Chih-En Kuo, Yi-Che Liu, Da-Wei Chang, Chung-Ping Young, Fu-Zen Shaw, and Sheng-Fu Liang. 2017. Development and Evaluation of a Wearable Device for Sleep Quality Assessment. *IEEE Trans. Biomed. Engineering* 64, 7 (2017), 1547–1557.
- Nicholas D. Lane, Mashfiqui Mohammad, and et al. 2011. Bewell: A smartphone application to monitor, model and promote wellbeing. In *Pervasive Computing Technologies for Healthcare*.
- Wen-Hung Liao and Chien-Ming Yang. 2008. Video-based activity and movement pattern analysis in overnight sleep studies. In *ICPR*. 1–4.
- Zhicheng Liao and Yangyong Zhu. 2014. When a Classifier Meets More Data. *Procedia Computer Science* 30 (2014), 50 – 59. 1st International Conference on Data Science, ICDS 2014.
- J.J. Liu, Wenyao Xu, and et al. 2013. A dense pressure sensitive bedsheet design for unobtrusive sleep posture monitoring. In *PerCom*. 207–215.
- Steven W Lockley, Debra J Skene, and Josephine Arendt. 1999. Comparison between subjective and actigraphic measurement of sleep and sleep rhythms. *Journal of sleep research* 8, 3 (1999), 175–183.
- Janna Mantua, Nickolas Gravel, and Rebecca Spencer. 2016. Reliability of Sleep Measures from Four Personal Health Monitoring Devices Compared to Research-Based Actigraphy and Polysomnography. *Sensors* 16, 5 (2016), 646.
- Shun Matsui, Tsutomu Terada, and Masahiko Tsukamoto. 2017. Smart eye mask: eye-mask shaped sleep monitoring device. In *Adjunct Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers, UbiComp/ISWC 2017, Maui, HI, USA, September 11-15, 2017*. 265–268.
- Mimo. 2016. <http://mimobaby.com/>. (2016).
- K. Nakajima, Y. Matsumoto, and T. Tamura. 2000. A monitor for posture changes and respiration in bed using real time image sequence analysis.. In *Engineering in Medicine and Biology Society*, Vol. 1.
- Yunyoung Nam, Yeesock Kim, and Jinseok Lee. 2016a. Sleep Monitoring Based on a Tri-Axial Accelerometer and a Pressure Sensor. *Sensors* 16, 5 (2016), 750.
- Yunyoung Nam, Yeesock Kim, and Jinseok Lee. 2016b. Sleep Monitoring Based on a Tri-Axial Accelerometer and a Pressure Sensor. *Sensors* 16, 5 (2016).
- Anh Nguyen, Raghdha Alqurashi, Zohreh Raghebi, Farnoush Banaei-kashani, Ann C Halbower, and Tam Vu. 2016. A Lightweight And Inexpensive In-ear Sensing System For Automatic Whole-night Sleep Stage Monitoring. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*. ACM, 230–244.
- Oakley NR. 1997. Validation with polysomnography of the sleep-watch sleep/wake scoring algorithm used by the actiwatch activity monitoring system.. In *Technical report to Mini Mitter Co., Inc*.
- Abdulakeem Odunmbaku, Amir-Mohammad Rahmani, Pasi Liljeberg, and Hannu Tenhunen. 2016. *Elderly Monitoring System with Sleep and Fall Detector*. Springer International Publishing, Cham, 473–480. https://doi.org/10.1007/978-3-319-47063-4_51
- James M. Parish. 2009. SLEep-related problems in common medical conditions. *CHEST Journal* 135, 2 (2009), 563–572.
- Rachael Purta, Stephen Mattingly, Lixing Song, Omar Lizardo, David Hachen, Christian Poellabauer, and Aaron Striegel. 2016. Experiences measuring sleep and physical activity patterns across a large college cohort with fitbits. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers, ISWC 2016, Heidelberg, Germany, September 12-16, 2016*. 28–35.
- Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning From Crowds. *J. Mach. Learn. Res.* 11 (Aug. 2010), 1297–1322.
- Yanzhi Ren, Chen Wang, Jie Yang, and Yingying Chen. 2015. Fine-grained sleep monitoring: Hearing your breathing with smartphones. In *2015 IEEE Conference on Computer Communications, INFOCOM 2015, Kowloon, Hong Kong, April 26 - May*

1, 2015. 1194–1202.

- Mahsan Rofouei, Mike Sinclair, Ray Bittner, Tom Blank, Nick Saw, Gerald DeJean, and Jeff Heffron. 2011. A Non-invasive Wearable Neck-Cuff System for Real-Time Sleep Monitoring. In *BSN. IEEE*, 156–161.
- A. Sadeh. 2011. The role and validity of actigraphy in sleep medicine: an update. *Sleep Med Rev* 15, 4 (Aug 2011), 259–267.
- A. Sadeh, K. M. Sharkey, and M. A. Carskadon. 1994. Activity-based sleep-wake identification: an empirical test of methodological issues. *Sleep* 17, 3 (Apr 1994), 201–207.
- Sohrab Saeb, Thaddeus R Cybulski, Konrad P Kording, and David C Mohr. 2017. Scalable Passive Sleep Monitoring Using Mobile Phones: Opportunities and Obstacles. *Journal of Medical Internet Research* 19, 4 (2017), e118.
- Alireza Sahami Shirazi, James Clawson, Yashar Hassanpour, Mohammad J. Tourian, Albrecht Schmidt, Ed H. Chi, Marko Borazio, and Kristof Van Laerhoven. 2013. Already Up? Using Mobile Phones to Track & Share Sleep Behavior. *Int. J. Hum.-Comput. Stud.* 71, 9 (Sept. 2013), 878–888.
- Aarti Sathyanarayana, Ferda Ofli, Luis Fernandes-Luque, Jaideep Srivastava, Ahmed K. Elmagarmid, Teresa Arora, and Shahrad Taheri. 2016. Robust Automated Human Activity Recognition and its Application to Sleep Research. *CoRR* abs/1607.04867 (2016).
- Aarti Sathyanarayana, Jaideep Srivastava, and Luis Fernández-Luque. 2017. The Science of Sweet Dreams: Predicting Sleep Efficiency from Wearable Device Data. *IEEE Computer* 50, 3 (2017), 30–38.
- Burr Settles. 2012. *Active Learning*. Morgan & Claypool Publishers.
- Sleep As Android. 2010. <https://play.google.com/store/apps/details?id=com.urbandroid.sleep&hl=en>. (2010).
- Sleep Bot. 2017. <http://mysleepbot.com>. (2017).
- Sleep Time Smart Alarm Clock. 2015. <https://play.google.com/store/apps/details?id=com.azumio.android.sleepertime>. (2015).
- Margarita Sordo and Qing Zeng. 2005. On Sample Size and Classification Accuracy: A Performance Comparison. In *Proceedings of the 6th International Conference on Biological and Medical Data Analysis (ISBMDA'05)*. 193–201.
- Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart Devices Are Different: Assessing and Mitigating Mobile Sensing Heterogeneities for Activity Recognition. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys '15)*. 127–140.
- Xiao Sun, Li Qiu, Yibo Wu, Yeming Tang, and Guohong Cao. 2017. SleepMonitor: Monitoring Respiratory Rate and Body Position During Sleep Using Smartwatch. *IMWUT* 1, 3 (2017), 104:1–104:22.
- Kristof Van Laerhoven, Marko Borazio, David Kilian, and Bernt Schiele. 2008. Sustained Logging and Discrimination of Sleep Postures with Low-level, Wrist-worn Sensors. In *Proceedings of the 2008 12th IEEE International Symposium on Wearable Computers (ISWC '08)*. 69–76.
- Oana Ramona Velicu, Natividad Martínez Madrid, and Ralf Seepold. 2016. Experimental sleep phases monitoring. In *Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on. IEEE*, 625–628.
- Vowpal Wabbit. 2016. <http://hunch.net/~vw/>. (2016).
- J B Webster, D F Kripke, and et al. 1982. An activity-based sleep monitor system for ambulatory use. In *Sleep*, Vol. 5. 389–399.
- HYERYEON YI, KYUNGRIM SHIN, and CHOL SHIN. 2006. Development of the Sleep Quality Scale. *Journal of Sleep Research* 15, 3 (2006), 309–316. <https://doi.org/10.1111/j.1365-2869.2006.00544.x>
- Zeo. 2003. <http://www.myzeo.com/sleep>. (2003).
- Jin Zhang, Qian Zhang, Yuanpeng Wang, and Chen Qiu. 2013. A real-time auto-adjustable smart pillow system for sleep apnea detection and treatment. In *Information Processing in Sensor Networks (IPSN), 2013 ACM/IEEE International Conference on*. 179–190. <https://doi.org/10.1109/IPSN.2013.6917584>