Detecting Drinking Episodes in Young Adults Using Smartphone-based Sensors

S. BAE, Carnegie Mellon University, Human-Computer Interaction Institute
D. FERREIRA, University of Oulu, Center for Ubiquitous Computing
B. SUFFOLETTO, J. C. PUYANA, and R. KURTZ, University of Pittsburgh, Department of Emergency Medicine
T. CHUNG, University of Pittsburgh, Department of Psychiatry
A. K. DEY, Carnegie Mellon University, Human-Computer Interaction Institute

Alcohol use in young adults is common, with high rates of morbidity and mortality largely due to periodic, heavy drinking episodes (HDEs). Behavioral interventions delivered through electronic communication modalities (*e.g.*, text messaging) can reduce the frequency of HDEs in young adults, but effects are small. One way to amplify these effects is to deliver support materials proximal to drinking occasions, but this requires knowledge of when they will occur. Mobile phones have built-in sensors that can potentially be useful in monitoring behavioral patterns associated with the initiation of drinking occasions. The objective of our work is to explore the detection of daily-life behavioral markers using mobile phone sensors and their utility in identifying drinking occasions. We utilized data from 30 young adults aged 21-28 with past hazardous drinking and collected mobile phone sensor data and daily Experience Sampling Method (ESM) of drinking for 28 consecutive days. We built a machine learning-based model that is 96.6% accurate at identifying non-drinking, drinking and heavy drinking episodes. We highlight the most important features for detecting drinking episodes and identify the amount of historical data needed for accurate detection. Our results suggest that mobile phone sensors can be used for automated, continuous monitoring of at-risk populations to detect drinking episodes and support the delivery of timely interventions.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI) → Empirical studies in HCI**

General Terms: Design, Algorithms, Detection

Additional Key Words and Phrases: Alcohol consumption, Smartphone sensors, Experience Sampling Method (ESM), Behavioral model, Machine learning, Young adults

ACM Reference Format:

Sangwon Bae, Denzil Ferreira, Brian Suffoletto, Juan C. Puyana, Ryan Kurtz, Tammy Chung and Anind K. Dey. 2017. Detecting Drinking Episodes in Young Adults Using Smartphone-based Sensors. *PACM Interact. Mob. Wearable Ubiquitous Technol.* X, X, Article X (Month YYYY), X pages.
DOI: 10.1145/1234

1 INTRODUCTION

Young adults have the highest prevalence (41%) of hazardous alcohol use among all age groups [11]. Young adults also report high rates of alcohol-related problems, including significant co-morbidities; unintentional injuries such as motor vehicle accidents; sexually transmitted infections; loss of productivity; broken relationships; and effects on physical health [1,46]. Unfortunately, numerous barriers prevent young adults from seeking help to reduce drinking [44] and rates of hazardous drinking among young adults have remained relatively unchanged in recent years [17].

Existing methods to detect a drinking occasion include self-reports, breathalyzer or transdermal alcohol monitors (*e.g.*, SCRAM ankle bracelet, WrisTAS) [38]. Self-reports of alcohol use have shown validity in specific contexts [53]. The use of Experience Sampling Methods (ESM) to collect self-reports of alcohol use more proximal to drinking occasions can minimize biases associated with retrospective reporting [38]. Self-report of drinking episodes using ESM is generally reliable and valid [54]. When comparing a breathalyzer with self-report of alcohol use (*i.e.*, start/end of drinking episode, quantity consumed) to estimate blood alcohol content, there is a high correlation [9], further supporting the validity and applicability of self-reported alcohol use. In comparison with SCRAM devices, self-reports produced a greater number of drinking events [4,26]. Moreover, SCRAM is sometimes subject to equipment failure (<10% of the time), less useful in detecting low drinking quantities compared to self-reports of alcohol use, and carries an associated stigma related to wearing an ankle monitor [4]. Another limitation of existing transdermal alcohol sensors is that transdermal alcohol content readings lag behind consumption by up to several hours, making the devices less useful for applications requiring real-time data [34]. However, issues of participant burden and reporting

compliance associated with self-reports, highlight the importance of developing alternative low burden methods to detect alcohol use quickly and unobtrusively.

The extensive body of research on drinking behavior suggests that young adults often consume alcohol in social contexts [6] and some young adults may not be ready to change drinking behavior [67]. Consequently, numerous drinking intervention studies, which applied social strategies and motivational factors, demonstrated effects of self-efficacy [63], self-control [28] clear advice to change [15], and personalized feedback [41] on reducing alcohol use. However, few studies focus on how to accurately detect drinking behaviors in the moment, especially heavy drinking in young adults in daily life. Young adults may not recognize when they are at risk of consequential harm from drinking excessively.

To support strategies for intervening in drinking behaviors, we construct a smartphone-based model to accurately detect drinking and heavy drinking episodes 'in the moment'. We can then use existing motivational strategies delivered in mobile messaging interventions [37] to help young adults to change behaviors *in the moment* and/or better reflect on drinking patterns and provide opportunities to regulate drinking patterns *after heavy drinking incidents* [61,62]. To extend this work, we envision a mobile message intervention, such that when drinking is detected, text messages that recommend the use of certain protective behavior strategies, such as *slowing the pace of drinking*, and *setting a limit on quantity to be consumed*, could be delivered when drinking is detected, to prevent a transition to heavy alcohol consumption. For example, a text message intervention that incorporated protective behavior strategies was effective in reducing binge drinking in young adults [61]. Similarly, a text message intervention focused on reducing weekend drinking was effective in reducing heavy drinking episodes up to 6 months post-intervention [62]. Young adults have expressed some willingness to receive intervention messages during drinking episodes [62]. Combining these results suggests that providing support (*e.g.,* suggestions for protective behavioral strategies) more proximal to drinking occasions, when drinking is accurately detected, could increase the likelihood that protective behavioral strategies will actually be used, thus reducing the frequency of binge drinking. We believe that detecting 'heavy drinking' could also be quite beneficial, particularly for those who frequently binge drink. Detecting drinking episodes in the moment would allow for a message to be sent to designated individuals (*e.g.*, designated sober driver) who could provide assistance and support. Also, with expert guidance and detected evidence of heavy drinking, individuals can then reflect on their drinking behavior, and gain insights into their drinking patterns; clinicians could use the data to adjust a care plan to address issues associated with heavy alcohol use.

Recently, a novel method for detecting substance use behavior was developed that applies machine learning-based models to sensor data, to continuously monitor physiological and behavioral patterns [20,49]. However, wearable sensor-based physiological monitoring can be burdensome and is not yet scalable, motivating our investigation of data available from built-in mobile phone sensors and meta-data (*e.g.*, call/text activity) to detect alcohol use occasions from daily activities. As an example of the potential utility of this approach, GPS digital activity trails over 4-5 hours predicted self-reports of heroin craving in substance users [49]. We hypothesize that using only smartphone-based sensor data, we can identify behavior patterns *(e.g.*, communication and travel patterns) that are associated with drinking occasions. For example, smartphone sensors could capture a behavioral pattern of increasing social activity within a specific time frame, indicated, for example, by increased texting and travel activity to meet with friends (*e.g.*, at a party or bar), followed by alcohol use [14]. Specifically, we examine the extent to which phone sensor data (*e.g.*, movement and motion sensor, geo-location, call and text patterns, and smartphone usage) contains features that could be useful in detecting drinking occasions.

Our focus in this study is to develop a machine learning model which detects not-drinking, drinking, and heavy drinking episodes using only data from smartphone-based sensors (*e.g.*, accelerometer, location) and meta-data (*e.g.*, communication logs) that capture individuals' daily activities. A primary potential application of this model is to support just-in-time delivery of intervention messages more proximal to drinking occasions, when preventive messages may be most salient and useful [57]. The model is not being developed to directly detect or estimate blood alcohol concentration using phone sensor data (*e.g.*, for legal or forensic purposes), which is beyond the scope of this study.

We used a mobile phone application, based on AWARE [24], to collect sensor data from smartphones and an automated text message program to collect self-reports of drinking events (*i.e.*, start/end of episode, number of drinks consumed) each day, in the morning, to increase compliance and minimize retrospective recall bias in responding to surveys sent directly to a personal smartphone. We identified the sensor-based features with the strongest relationships to *not drinking*, *drinking* and *heavy drinking* episodes, and determined the accuracy of machine-learning based models using these features to detect and differentiate not drinking, drinking and heavy drinking episodes.

In this paper, we provide two main contributions. First, using only smartphone data collected from 30 young adult participants, we built a machine learning based-model that can detect whether an individual is not drinking, drinking, or heavy drinking based on self-reports in the natural environment with an accuracy of 96.6%. By comparison, prior work, which used the accelerometer from a smartphone (*e.g.*, gait) conducted in a lab setting, obtained only 70% accuracy in detecting the number of drinks consumed [2]. Our study provides an advance over existing research by using data collected in the natural environment (versus the lab), in a larger sample, and using data available only from a smartphone that most young adults already carry (rather than a wearable device that few people own). More importantly, we identify the most important features for performing this detection, which can be used

to inform the timing of intervention delivery. Second, we determine the relative value of using different amounts of historical data, and different size windows of data on detection accuracy to maximize the efficiency of the model. In this regard, we found that 1-day historical data on smartphones is enough to detect drinking and heavy drinking episodes. The former indicates how much data needs to be stored on the phone for accurate drinking detection, and thus, the privacy risk. The latter is informative for determining optimal time windows for intervention delivery.

In the following section, we review previous work related to developing smartphone-based behavior models. Next, we describe our method for data collection, feature selection, and detecting not drinking, drinking and heavy drinking behaviors using our machine-learning based model. We then conclude with a discussion of the implications and contributions of our model development.

2 BACKGROUND AND RELATED WORK

2.1 Smartphone-based behavior modeling

In the field of ubiquitous computing, smartphone instrumentation has enabled better understanding of users' interaction with these devices in specific contexts. For example, they have increased our understanding of how people use applications [3,22] and smartphone networks [23], and allowed us to predict which application is relevant to the current context [35,50], and to detect the most opportune moments to deliver information to users [32]. More related to our work, in the area of health and wellbeing, the widespread availability of smartphones in today's young adult population has prompted research that leverages the embedded sensors in smartphones to study human behavior.

Researchers have used smartphones to assess and predict academic performance [65], used them to detect sleep and sleep quality [36], and personality traits [13], to passively sense and detect mental health changes (*e.g.*, schizophrenia [64], lack of social interaction [18]), and to detect habitual behaviors such as smoking [52]. It is noteworthy that substance use (*e.g.*, cocaine usage [10], cigarette smoking [49], heroin craving [20]) can be detected using machine learning applied to data from wearable sensors. However, wearable sensors can be burdensome, and their use does yet not scale to long periods of time nor large numbers of users. This has motivated our work in understanding how the combination of machine learning and sensor data from commodity mobile phones can be used to detect drinking episodes, particularly to support delivery of messages for just-in-time and *post hoc* intervention [39,40].

2.2 Defining "Drinking" and "Heavy Drinking" episodes

The National Institute on Alcohol Abuse and Alcoholism [43] defines a standard drink as "any drink that contains about 14 grams of pure alcohol, i.e., 1.2 tablespoons." To illustrate the NIAAA's standard drink measurement, one 12 oz. beer is equal to 1 standard drink, one 16 oz. malt liquor is 2 standard drinks. One mixed drink with "hard liquor" is estimated – depending on the alcohol percentage – to contain one or more standard drinks, where a pint (16 oz.) of 80-proof alcohol is equivalent to 11 standard drinks. Moderate alcohol consumption is 2 and 1 standard drinks (men and women, respectively) per day. The NIAAA defines binge drinking as a pattern of drinking that brings blood alcohol concentration (BAC) levels to 0.08 g/dL [44]. This equates to 4 or more standard drinks for women or 5 or more standard drinks for men consumed in roughly 2 hours [44]. We perform our analysis with heavy drinking defined as any drinking occasion when an individual reported either consuming ≥4 drinks (for women) or ≥5 drinks (for men) [42].

2.3 Methods to Assess Drinking Behavior: Self-Report and Wearable sensors

Research on drinking behavior in the psychology, nursing and medical domains have traditionally quantified participants' alcohol consumption with self-reports (*e.g.*, interview, questionnaire, diaries, ESM), observer reports, and sensors (Table 1).

Table 1. **Drinking Behavior Studies and Methodology**

Paper	Year	Sample, Length	Research question	Methods & tools	Analysis	Accuracy	Limitations & Future work
[27]	2016	N=5, 1d	Smartwatch-based user's alcohol intoxication level estimation	Sensor instrumentation	Regression SVM ANN	32% 88,6% 52,4%	Laboratory study
[54]	2015	N=60, 2y	Examine the convergent validity of three approaches to collecting	Biochemical (WrisTAS), daily and weekly	Multilevel logit model (Stata 13)	85,7%	Limited number of participants, and assessment days;

			daily self-report drinking data: experience sampling, daily morning reports of the previous night, and 1-week timeline follow-back (TLFB) assessments.	reconstruction			larger sample and replication is needed
[2]	2015	N=6, 2w	Smartphone-based user's alcohol intoxication level (how many drinks) can be inferred from their gait, in three categories: 0-2 as sober, 3-6 as tipsy and >6 drinks as intoxicated.	Sensor instrumentation, self-report	Random Forest	56,0%	Very limited sample and evaluation; future work should include additional sensors: gyroscope, GPS, BT, magnetometer and other inertial sensors.
[50]	2015	N=213, 1y	Assess the role of personality and drinking onset in predicting weekly alcohol consumption, and the impact of the whole set of variables in predicting the number of consequences associated with consumption in undergraduates.	Weekly reconstruction	Hierarchical regression	20,2%	Participants have different notions of how to measure binge drinking
[16]	2014	N=312, 1y	Test a model including Facebook alcohol displays and constructs from the theory of reasoned action to predict binge drinking.	Social media	Path modeling	NA	Sample not representative of all colleges, ethnically and racially diverse; self-report bias on recall and social desirability bias.
[21]	2013	N=37, 16m	Study the longitudinal effects of alcoholism on gait and balance	Interview, observations	Statistical analysis	NA	Small sample, found no evidence of change after 1 year, unsure why
[19]	2012	N=44,610, 1y	Analyze ultimate and distal factors predicting substance consumption according to Petraitis' theory of triadic influence	Survey	Multivariate logistic regression	70,2%	Limited to Germany, adolescents
[45]	2011	N=81, 1d	Compare trained field observer reports of number of drinks consumed with participant self-report (1-2 days after drinking episode) of drinking quantity	Next day self-report collected by phone interview	Correlation, comparison of means	NA	Participants accurately estimated their consumption when consuming eight or fewer drinks in a single session; underestimated consumption above eight drinks
[55]	2010	N=423, 1y	Evaluate the impact of online social-norms	Survey	Generalized linear mixed models	NA	Intervention used physical methods

			interventions (personalized, marketing ads, attention control)				(flyers, banners, campus newspapers), uncertain of reach, impersonal; limited data
[5]	2008	N=52, 2w	Explore the relationship between alcohol consumption (i.e., binge) and health	Survey	Descriptive statistics	NA	Short term; limited data
[42]	2007	N=818, 1y	Evaluate the relative contribution of social norms, demographics, drinking motives, alcohol expectancies in predicting alcohol consumption	Survey, weekly reconstruction	Regression	24,0%	Focused on first year students, may not apply to senior
[14]	2003	N=1909, 2y	Examine the validity of a set of environmental variables to predict heavy drinking at college students' most recent drinking occasions	Interview, survey	Nonparametric exploratory and confirmatory discriminant analysis	48,0%	Limited data, future studies would benefit from inclusion of more indicators of the social and physical environment in which college students drink
[66]	2003	N=1894, 4m	Identify person, social group, and environmental factors associated with uptake of binge drinking among national college students	Survey	Univariate and multivariate logistic regression, Generalized Estimating Equations (GEE) from Statistical Analysis Software	NA	Focused on first year students; limited contextual data
[48]	1999	N=3961, 4y	Compare three methods for assessing alcohol consumption to resulting prevalence estimates for high risk drinking and harm as defined by morbidity and mortality indicators	Survey, weekly reconstruction	Cross-tabulation, Spearman correlation, and descriptive statistics	NA	Based on estimated values; limited contextual data

However, retrospective reports, especially over long intervals (*e.g.*, past month) may be subject to bias. ESM methods (*e.g.*, report of prior day's drinking behavior) generally provide higher rates of self-reported drinking episodes compared to methods that ask individuals to recollect drinking behavior over longer periods (*e.g.*, past month) [51].

2.4 ESM serves as "ground truth" for drinking behavior

The main methods of measuring alcohol use include self-report, and breath or transdermal alcohol monitors. Daily self-report of drinking may be subject to underreporting, but given proper assurances and appropriate data collection methods (*e.g.*, ESM), as used in this study, participants can provide reliable and valid self-report of alcohol use [54]. We used a daily morning ESM report to obtain data on the prior day's alcohol use, because research comparing random ESM, end of the day ESM, and next day ESM report found that next day ESM in the morning provided a better summary of the prior day's drinking (*i.e.*, more drinking events reported, and higher quantity of alcohol consumed per day) than other self-report methods [54]. In addition, a preliminary study that we conducted resulted in relatively low completion (32%) rates for hourly reports of alcohol consumption (*e.g.*, push notifications sent every hour from 8pm to 12am on weekend days after onset of a drinking episode was reported). Exit interviews indicated that participants either ignored or did not remember to complete ESM self-reports during or at the end of the drinking

episode. For these reasons, we used daily morning ESM reports to obtain data on the prior day's alcohol use to minimize participant burden and reduce potential retrospective recall bias.

Studies comparing ESM report of alcohol consumption against WrisTAS and SCRAM generally find that self-reports generate more drinking days than WrisTAS or SCRAM [4,26,54]. Specifically, a study that compared multiple forms of self-report (*e.g.*, recall of past week, ESM) with transdermal alcohol bracelet (WrisTAS) found that ESM corroborated with 85.74% (biochemical) and 87.27% (daily morning reports of previous night and 1-week timeline follow-back) of drinking days [45]. Other work found that the sensitivity of the SCRAM device (ankle monitor) exceeded WrisTAS in detecting self-reported drinking events [26]. A review of SCRAM studies indicated that the SCRAM detected 73%-91% of self-reported drinking days [8]. SCRAM sensors showed good ability to detect >5 drinks, but appeared to be less sensitive at lower drinking quantities [4]. Use of transdermal alcohol sensors also needs to consider the time lag between transdermal alcohol content and breath alcohol concentration, which averaged over 2 hours (129 minutes) [29], limiting the utility of these sensors for real-time detection of alcohol use. Studies that have collected self-report (using ESM, web diary) on start/end time of drinking, and number of drinks consumed, in order to compute estimated BAC find that self-report provides estimates of BAC that are strongly correlated with breathalyzer readings [9]. In sum, research comparing WrisTAS, SCRAM, and breathalyzer with self-report of alcohol use indicate that self-reports produced more data on drinking days compared to WristTAS and SCRAM, and had strong correspondence with breathalyzer reading, supporting the use of ESM self-report to assess alcohol use in this study.

Based on the literature and our preliminary work, we used ESM self-reported data on alcohol consumption as "ground truth" in developing a model to detect drinking episodes based on smartphone sensor data. We, therefore, leverage widely used smartphones and their sensing ability to detect instances of drinking in the wild, and to perform daily experience sampling. In addition, looking ahead to the future, the smartphone can also be used as an intervention delivery platform. In previous work, we found that interactive text messages were successful in reducing heavy drinking episodes among young adults [58-61].

3 METHOD

3.1 Participants

A convenience sample of Emergency Department (ED) patients aged 21 to 28 years were identified over three months. Those medically stable and not seeking treatment for substance use disorder were screened for enrollment. We included 21 individuals who reported recent hazardous alcohol consumption based on an Alcohol Use Disorder Identification Test for Consumption (AUDIT-C) score of ≥3 for women or ≥4 for men [7] and at least one heavy drinking episode (HDE) on any day in the prior month.

We also recruited 17 young adults (ages 21-28 years) from the Craigslist website, a local participant pool and through study flyers placed on and off campus at locations such as the university student center and a nearby coffee shop. Thirteen of these individuals were included in our study, using the same screening as above (AUDIT-C and HDE report). The individuals represented undergraduates, graduate students, and young professionals who regularly used a mobile device (Android or iPhone).

In total, 38 (21 from ED, 17 from general population) young adults (50% female, mean age=23.15, SD=1.89) met our enrollment criteria, provided informed consent, answered a questionnaire (with questions about demographics, height and weight) and downloaded our data collection app to their phone. Participants were compensated $20 for completing the baseline questionnaire and installing the smartphone app. For each day on which an ESM was completed they also earned $2, which was paid at the end of the 28-day study.

3.2 Data Collection

The AWARE-based application passively collected the timestamped sensor data shown in Table 2. For clarity, 1 Hz is 1 sample per second.

Table 2. **Mobile Phone Sensors and Frequency of Collection**

Sensor	Frequency
Accelerometer: to detect motion and device interaction	20Hz
Keyboard: i.e., keystroke speed, text input length.	Event-based
Battery usage: battery level in percent and voltage, charging state, battery temperature	Event-based
Communication: meta-data from calls and texts, i.e., timestamp, one-way hashed phone number, and if received, sent, or missed	Event-based
Device usage: amount of time the device is in use and when idle	Event-based

Mobility State: still, tilting, walking, running, on bicycle, in vehicle, using an Activity Recognition API	1-minute interval
Luminosity: to detect well-lit environments	1 Hz
Network usage; traffic: WiFi, Bluetooth, airplane on/off states; and bytes and packets exchanged	Event-based; 10-second interval
Location: using the phone's fused location provider – best location estimate (~50 meters) using simultaneously cell towers' positioning, GPS and WiFi access points	3-minute interval
Proximity to screen: to detect if the phone is in one's pocket	1 Hz
Rotation: to detect holding of the device	20Hz
Screen status: on, off, locked, unlocked	Event-based
Telephony: connected cell tower ID and nearby towers, with signal strength	Event-based
WiFi: nearby Wi-Fi access points and signal strength	1-minute interval

When installing AWARE for the first time, a Universal Unique ID (UUID) is randomly generated. We use this UUID identifier to identify a participant without storing any personal data (*e.g.*, name). The application initially stored the sensor data from the table above on a participant's device and then synchronized this information to our server over a secure connection (SHA-256, 2048-bit long RSA encryption key), via Wi-Fi only. A sync attempt is performed every 30 minutes. If the user is not connected to Wi-Fi, the synching is postponed for another 30 minutes.

GPS has been used previously to monitor one's travel activities and interactions with the smartphone for mental health applications [20,64]. However, to the authors' best knowledge, no work exists using sensors to infer social and behavioral context in association with drinking episodes. Moreover, we hypothesize that communication activities such as call and messaging events collected by a smartphone could be used to detect drinking events. For example, young adults may increase communication activities with friends or colleagues just prior to drinking events to plan and decide for the evening's activities at a party or bar.

To identify drinking episodes and to calculate an estimate of blood alcohol content, we triggered a notification for a survey each day at 10am for 28 days, asking participants the following survey questions:

"Did you drink alcohol yesterday?"

If they responded "no" then no further questions were asked. If they responded "yes", we asked:

"Approximately what time did you start drinking?",

"Approximately what time did you stop drinking?", and

"How many standard drinks did you have during this period?".

Participants were provided with the definition of a standard drink (*e.g.*, 12 oz. can of beer or 5 oz. glass of wine or 1.5 oz. 80-proof liquor) at the start of the study [43]. We trained participants on the definition of a standard drink, and sent an illustration (in the survey) of a typical standard drink for common beverage types: beer, wine, liquor. If there were multiple drinking episodes in a single day, participants were instructed to report on the occasion when the largest number of drinks was consumed.

3.3 Measurement of drinking behavior

Ground truth for our analyses was self-reported alcohol use based on daily surveys via smartphones. We defined a drinking occasion as any day in the 28-day data collection period when an individual reported "yes" to the question "Did you drink alcohol yesterday?" A "heavy drinking episode" was defined as any drinking occasion when an individual reported either consuming ≥4 drinks (for women) or ≥5 drinks (for men) [42]. A "drinking episode" was any episode where alcohol was consumed that did not meet the heavy drinking episode criterion.

There are three points at which accuracy can come into question when using self-reports: remembering to report a drinking episode, remembering the number of drinks in the drinking episode, and remembering the timing (start and end) of the drinking episode. Based on the existing literature (section 2.4), we used ESMs to obtain self-report information on drinking episodes.

3.3.1 Reporting a drinking episode

The existing literature shows that self-reports generally produce a greater number of drinking events compared to episodes extracted from the SCRAM or WrisTAS devices [4,26]. Self-reported drinking episodes using ESM also have been shown to be generally reliable and valid [54]. Further, alcohol sensors such as SCRAM were sometimes subject to equipment failure (<10% of the time), and were less useful in detecting low drinking quantities compared to self-report of alcohol use (Barnett *et al*., 2014).

3.3.2 Reporting the number of drinks

Research indicates that self-reported (next day) number of drinks and the time of drinking onset are highly correlated (0.84) with data from a breathalyzer test (breath alcohol concentration or BrAC) [9]. In addition, a more recent study [4] showed that the transdermal alcohol concentration from SCRAM was highly correlated with self-reported number of drinks ($r<0.77$, $p<0.001$). Another line of evidence involves a study that compared trained field observer reports of the number of drinks consumed in the natural environment (*e.g.*, bar) with an individual's self-report (obtained by phone interview 1-2 days after the drinking episode) [45]. Compared to trained observers, participants accurately estimated their consumption, particularly when consuming eight or fewer drinks in a session [45], a quantity that is lower than that used to define a "heavy drinking episode" for our current analyses.

3.3.3 Reporting the timing of a drinking episode

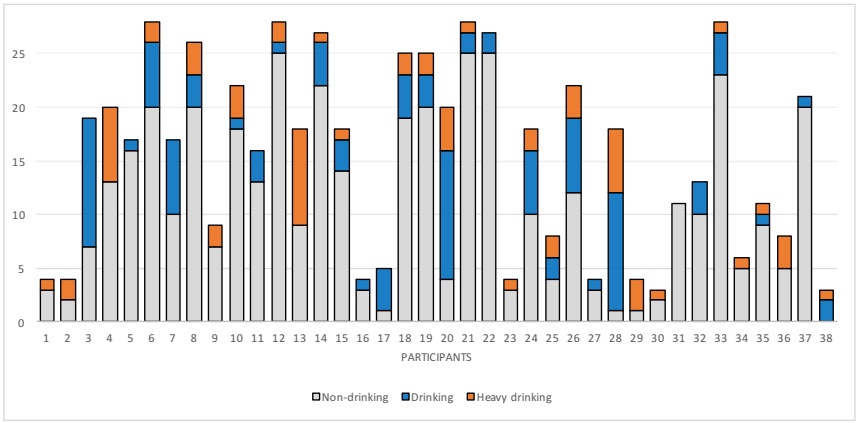
Self-reported onset and ending of drinking events, combined with self-reported number of drinks consumed, was strongly correlated with breathalyzer readings [9], providing some support for the validity of self-reported times for start/end of drinking. By contrast, readings of alcohol use from devices such as SCRAM and WrisTAS involve a time lag in detection of alcohol by up to several hours, making the devices less useful for applications needing real-time data [34]. SCRAM and WrisTAS are also limited with regard to determining drinking start time due to their reduced sensitivity in detecting lower levels of alcohol use (*e.g.*, <5 drinks), which occur at the start of drinking episodes [26]. In general, self-reports of start time for a drinking episode likely provide greater precision relative to existing transdermal sensors (Dr. Denis McCarthy, Director of the Alcohol Cognitions Lab at the University of Missouri, personal communication: January 2017). Overall, the literature supports the accuracy of self-reported drinking start time.

3.4 Drinking Detection Model Development

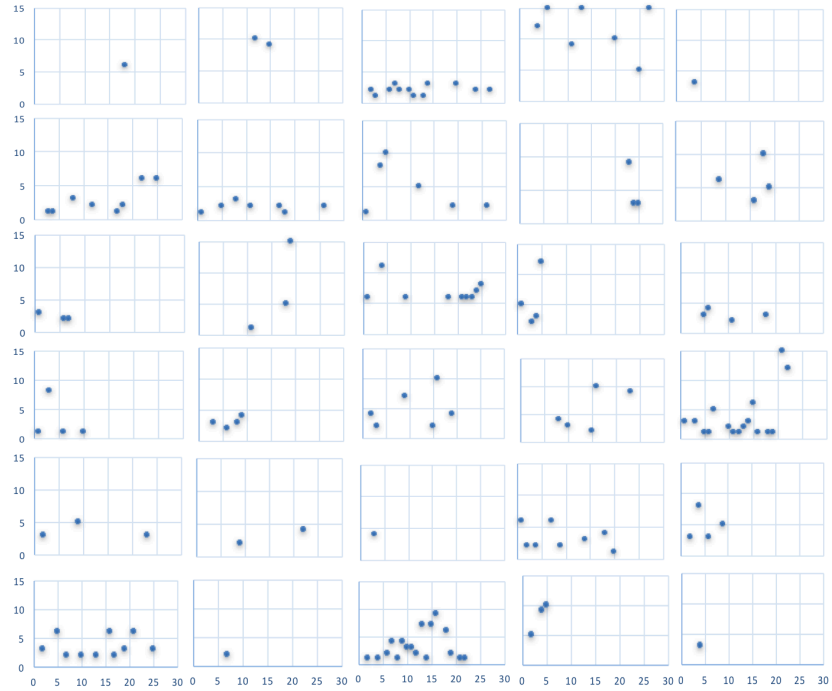
Developing a model for detecting drinking involved a four-step process: data pre-processing, data preparation, feature extraction, and training of classification models.

3.4.1 Data Pre-processing

Our 38 participants reported 621 episodes (Fig. 1, a); 415, 135 and 71 for non-drinking (M=10.92, SD=7.92), drinking (M=3.55, SD=4.32) and heavy drinking (M=1.86, SD=2.19), respectively. Our first inclusion criterion was self-report of at least 1 non-drinking episode and at least 1 drinking or heavy drinking episode. We excluded 2 subjects because one reported only 11 non-drinking episodes and no other episodes, and the other reported only 21 drinking and 7 heavy drinking episodes but did not report any non-drinking episodes. Our second inclusion criterion was that participants had to keep their smartphones on and not disable any of the sensors. We excluded an additional six participants who had only had one or two days' worth of sensor data (corresponding to 72 non-drinking, 14 drinking and 5 heavy drinking episodes). They did not have sufficiently granular sensor data (*i.e.*, they manually disabled location or motion sensor plug-in, or explicitly turned off the smartphone for long periods of time). The remaining 30 participants had 332 non-drinking, 100 drinking and 59 heavy drinking episodes (Fig. 1, b). Our final inclusion criterion was that we needed to have sensor data for those drinking episodes. We excluded episodes if we were missing sensor data for them. The focus of our analysis was the remaining 293 episodes from 30 participants: heavy drinking (45) or drinking (41) reports, and non-drinking episodes (207) because 125, 59 and 14 for non-drinking, drinking and heavy drinking episodes were removed due to lack of sensor data.



(a) ESM reports across the participants (n=38) including excluded participants (31-38)



(b) Drinking reports per each participant (n=30): x-axis refers to days during the study (max=28 days), y-axis refers to the number of standard alcohol drinks consumed

Fig. 1. **ESM reports of drinking episodes for each participant.**

If there were missing sensor values at a certain timestamp, we interpolated the average value between two instances rather than simply removing data. In addition, we used the day of week as nominal attributes (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday).

3.4.2 Data Preparation

All participants (without prompting) reported the start and end times of their drinking episodes with a granularity of 15 minutes (*e.g.*, 7:00, 7:15, 7:30, 7:45), likely due to the fact that self-reports were not provided until the next day. With this self-report granularity, we chose to use 5-minute windows as our analysis window. We divided the time-series sensor data from each participant into a series of non-overlapping 5-minute windows and utilized the 5-minute window as our base unit for analysis and for extracting features. For the 30-minute, 1-hour and 2-hour windows we took the average of the numerical sensor values, and the most frequent amongst the nominal sensor values, among six, twelve, and twenty-four segments, respectively, of the base unit. The time of day was coded as 1 to 24 if dataset was 1-hour segment instances. 1 starts from 00:00 to 00:59 and 24 refers to the time between 23:00 to 23:59.

If a participant reported not drinking during the previous day in the survey, we labeled all of data windows for that day as *non-drinking* (N). When they did report drinking episodes, we labeled the windows before the start time and after the end time as *non-drinking* (Fig. 2 – top). For the windows during the reported drinking episode, we labeled them as *drinking* (D) if the number of drinks consumed was less than 4 (for female participants), or 5 (for male participants). Otherwise, they were labeled as *heavy drinking* (H).

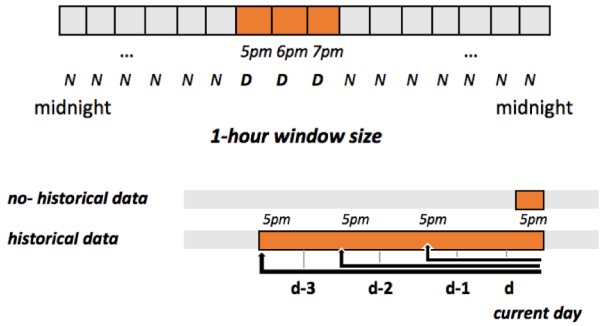


Fig. 2. **Description of window size (top) and use of historical data (bottom).**

We believe that social and behavioral data captured by smartphone sensors can help to detect the current drinking episodes [14]. Instead of only looking at a single time window for extracting features and training drinking detection models, we also considered the use of *historical data*, *i.e.*, data that preceded the drinking episodes. As shown in Fig. 2 – bottom, if participants reported a heavy drinking episode that started at 5pm, then we considered sensor data from 5pm on the previous day up to 5pm on the drinking day as a 1-day historical dataset. In our analyses, we considered 1-day, 2-day and 3-day histories referring to the social and behavioral sensor streams that were captured and stored on smartphones before drinking episodes began.

The final dataset included 12,442 segments (11,798 non-drinking, 243 drinking and 401 heavy drinking). We split our data into a training dataset (60% of all episodes, on which to build models), a cross-validation set (20% of all episodes to optimize our models), and a testing dataset (20% of all episodes, on which to test our model) (Table 3).

Table 3. **Dataset Distribution**

Models	Three Classes	All	Training set (60%)	CV set (20%)	Test set (20%)	SMOTE Training set 800% for Drinking & 400% for Heavy drinking (oversampled).
30min window size -no history data	Non-drinking	11798	7078	2360	2360	7078
	Drinking	243	145	49	49	1160
	Heavy drinking	401	240	80	81	960
30min window size -1day historical data	Non-drinking	2652	1591	531	530	1591
	Drinking	204	122	41	41	976
	Heavy drinking	388	233	78	77	932
30min window size	Non-drinking	4510	2706	1127	1128	2706

-2days historical data	Drinking	226	135	56	57	1080
	Heavy drinking	387	232	97	97	928
30min window -3days historical data	Non-drinking	6044	3626	1209	1209	3626
	Drinking	236	141	48	47	1128
	Heavy drinking	387	232	78	77	928
1hour window no- historical data	Non-drinking	5870	3522	1174	1174	3522
	Drinking	139	83	28	28	664
	Heavy drinking	214	128	43	43	512
1hour window size -1day historical data	Non-drinking	1437	862	288	287	862
	Drinking	101	60	21	20	480
	Heavy drinking	207	124	42	41	496
1hour window size -2days historical data	Non-drinking	2385	1431	477	477	1431
	Drinking	111	66	23	22	528
	Heavy drinking	207	124	42	41	496
1hour window size -3days historical data	Non-drinking	3137	1903	635	635	1903
	Drinking	117	70	23	24	560
	Heavy drinking	207	124	41	42	496
2hour window size no historical data	Non-drinking	2941	1764	588	580	1764
	Drinking	68	40	14	14	320
	Heavy drinking	104	62	21	21	248
2hour window size -1day historical data	Non-drinking	721	432	144	145	432
	Drinking	60	36	12	12	288
	Heavy drinking	100	60	20	20	240
2hour window size -2days historical data	Non-drinking	1260	756	252	252	756
	Drinking	65	39	13	13	312
	Heavy drinking	100	60	20	20	240
2hour window size -3days historical data	Non-drinking	1689	1013	338	338	1013
	Drinking	70	42	14	14	336
	Heavy drinking	100	60	20	20	240

To account for the imbalanced class sizes where heavy drinking and drinking episodes represent a minority of the data compared to non-drinking events in our dataset, we used SMOTE (Synthetic Minority Over-Sampling Technique) in the *training* set when we built our models [12]. This technique is a standard balancing approach that oversamples the instances of the underrepresented target event. Rather than just creating copies of these events, it selects two or more (*k*-nearest neighbor*)* similar instances using a distance measure and generates synthetic samples by perturbing attributes of one of the instances by a random amount, such that the similarity of the two instances remains within the original distance.

The technique can be also applied to a multiple class problem such as ours as well. As Table 3 shows, for each setting (window size x amount of historical data), there were several instances in the majority non-drinking class (N) and instances in the minority drinking class (D) and heavy drinking (H) respectively. We used this data to oversample (setting *k* to 5) the minority classes, drinking and heavy drinking classes at 800% and 400% of their original size, respectively, from the training set compared to models using original training set [12]. To determine the optimal oversampling, we calculated the ROC convex hull, a common approach for estimating the performance of classifiers for imbalanced datasets. Using this oversampling approach, we ended up with 7078, 1160 and 960 instances for non-drinking, drinking and heavy drinking classes respectively compared to the original 7078, 145 and 240 instances for a 30-minute window size when not using historical data (Table 3). Previous literature pointed out

that this approach efficiently leads the decision region of the minority class to become more general, making the classifier less specific but having bigger decision regions [12].

3.4.3 Feature extraction

With the resulting balanced data, we calculated features derived from both the "raw" smartphone data (*e.g.*, GPS, call, message, activity, accelerometer, and meta-data) and computed values (*e.g.*, {minimum, median, maximum, average and standard deviation}) of acceleration for each window size; 30-minute, 1- and 2-hours. First, statistical features of location (*e.g.*, travel distances and radius of gyration) within time segments were extracted to capture movement patterns. Second, we extracted the duration and number of incoming/outgoing messages, calls and contacts features to understand individuals' communications associated with drinking episodes. Third, physical activities and motions were extracted to capture human behaviors. Finally, we used screen on and off, battery status, charging time and length of charge to understand how people use smartphones during non-drinking and drinking episodes. In total, we extracted 56 features to represent the episodes (Table 4).

Table 4. **Extracted Features from each Sensor Stream**

Sensors	Contextual information derived
Time	Day of week
Location	Time at/away from home, number of places visited, travel distance, number of changes in location, time spent in a certain location, entropy, radius of gyration
Communication	Number of incoming and outgoing calls and text messages, number of contacts, duration of incoming and outgoing calls, number of incoming and outgoing calls
	Speed of typing, number of insert, delete, number of emojies, types of emojies, frequent timeslots of typing, number of conversations, length of conversation
Motion	{Min., Med., Max., Avg., Std.} of the changes in activity, number of activities per time slot, number of changes in activity
	{Min., Med., Max., Avg., Std.} Magnitude of acceleration
	{Min., Med., Max., Avg., Std.} Magnitude of rotation
Device usage	Name and type of applications, frequency and length of use, number of changes between applications, number of applications running
	Frequency and duration of screen on and off
	battery status, charging time, length of charge
	{Min., Med., Max., Avg., Std.} of screen proximities

3.4.4 Classifier building

To decide which features to include in our models of drinking detection, we took two approaches. First, we run the correlation analysis to gain an intuitive understanding of the value of individual features, and to see the value of the historical data in detecting non-drinking and drinking episodes. Second, we applied an attribute evaluator, Information Gain, and a ranker to identify the 20 most informative features. Information Gain is the amount of information that is gained by a model by knowing the value of a particular attribute or feature [31]. Using these top-20 features, shown in Table 6 and Figure 5 below, we trained the following machine learning classifiers: C4.5 decision tree, Bayesian Network (BN) and Random Forest (RF) used by Weka.

To compare different models' performances, as we mentioned, labeled episodes were divided into time windows by splitting into 10 groups and training on 9 selected groups, tested on the remaining group, and repeating this process 10 times, once for each of the 10 groups. We conducted 10-fold cross-validation on the training dataset (60% of data) of all labeled episodes across all users. We evaluated our resulting models using 10-fold cross-validation in the cross validation (CV) dataset (20%) and compared the models' performances.

The models' performances were evaluated using accuracy, F-score and ROC area under the curve (convex hulls) which are traditional methods for comparing machine learning model performance. Accuracy approximates how effective the algorithm is by showing the probability of the true value of the class label (assesses the overall effectiveness of the algorithm); ROC represents a relation between the sensitivity and the specificity of the algorithm; F-score is a composite measure which favors algorithms with

higher sensitivity and challenges those with higher specificity. We used the F-score to optimize parameter values for our three chosen machine learning algorithms (C4.5, BN and RF), and used the default settings in Weka.

In addition to the 3 different types of classifiers, we also experimented with different parameters: *different data window sizes and different amounts of historical data*. Our contribution here is two-fold. First, by experimenting with different amounts of data, we can provide guidance to other researchers about how much data is required for accurate detection and how much data needs to be stored on the phone impacting user privacy. Second, by experimenting with different data window sizes and identifying important features, we can help clinicians decide when to intervene, the types of data that might be useful to include in the intervention, respectively.

4 RESULTS

In this section, we describe the behavioral model we built to detect non-drinking, drinking and heavy drinking episodes among young adults using data continuously collected from their smartphone sensors during everyday life. In addition to reporting the accuracy of our model, we identified 1) which features significantly correlate to non-drinking, heavy drinking and drinking episodes, and suggest the most important features, 2) the impact of the time window size on model performance for detecting drinking episodes and 3) how much historical sensor data is needed to achieve the best-performing model. After we present the model, we will discuss how our model can facilitate interventions once drinking episodes are detected. We will now describe how we identify the most important features, with correlation analysis and information gain.

4.1 Correlation Analysis: Important Features of Drinking Episodes using Smartphone Sensor Data

We ran the correlation analysis 1) to gain an intuitive understanding of the value of individual features, 2) to quantify the strength of the relationship between sensor variables and not-drinking, drinking and heavy drinking episodes, and 3) to understand the value of using historical data for differentiating drinking episodes. In addition, we explain how the features differ when detecting non-drinking, drinking, and heavy drinking. For this analysis, we divided the data into 1-hour segments, providing enough data on which to operate, without overburdening the analysis with a data window that is too large.

To test if there is a linear relationship between the variables, we used the Pearson correlation coefficient r (from the Hmisc package of the R program to compute the significance levels for Pearson correlations). Table 5 shows the results of the correlation analysis where we present the features only if they have a positive or a negative relationship with whether a participant was not drinking, drinking, or heavy drinking. The correlation coefficient ranges from +1.0 to -1.0. $r > 0$, for example **time_of_day** $r = 0.11$ (Table 5a), indicates a positive linear relationship; $r < 0$, for example **accelerometer_mean_magnitude** $r = -0.03$ (Table 5a), indicates a negative linear relationship.

Table 5. **Correlation Matrix with Significance Levels (p-value)**

Features	r	p-value
time_of_day	0.11	0.000
screen_duration_interaction_seconds	0.07	0.000
day_of_week	0.06	0.000
average_time_between_keypress_ms	0.06	0.000
number_of_keypress_deletions	0.06	0.000
accelerometer_min_magnitude	-0.05	0.0001
number_of_keypress_insertions	0.04	0.002
accelerometer_mean_magnitude	-0.03	0.023
count_activity_changes	0.03	0.030
accelerometer_median_magnitude	-0.03	0.039
number_of_correspondents	0.03	0.047

(a) Correlations to episodes without historical data (n=6223)

Features	r	p-value
screen_unlocks_per_minute	0.20	0.000
time_of_day	0.19	0.000
screen_duration_interaction_seconds	0.15	0.000
number_of_deletions	0.15	0.000
happy_emoticon_count	0.14	0.000
day_of_week	0.10	0.000
average_time_between_keypress_ms	0.07	0.007

number_of_insertions	0.05	0.022

(b) Correlations including 1 day of historical data before drinking episodes (n=1745)

Features	r	p-value
time_of_day	0.16	0.000
happy_emoticon_count	0.11	0.000
day_of_week	0.08	0.000
screen_duration_interaction_seconds	0.08	0.000
average_time_between_keypress_ms	0.08	0.000
number_of_deletions	0.07	0.0006
number_of_insertions	0.06	0.002

(c) Correlations including 2 days of historical data before drinking episodes (n=2701)

Features	r	p-value
time_of_day	0.14	0.000
happy_emoticon_count	0.10	0.000
day_of_week	0.08	0.000
screen_duration_interaction_seconds	0.08	0.000
average_time_between_keypress_ms	0.07	0.0001
number_of_keypress_deletions	0.06	0.0002
number_of_keypress_insertions	0.05	0.003
radius_of_gyration	0.03	0.05

(d) Correlations including 3 days of historical data before drinking episodes (n=3494)

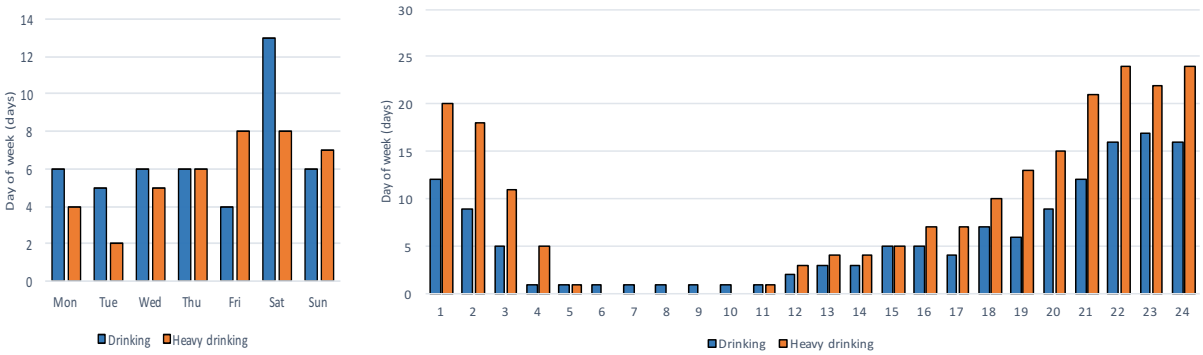
We performed the correlation analysis with historical data (1-day, 2-days, 3-days) and without historical data. Table 5a shows the results of the analysis with no historical data, and Table 5b, 5c and 5d shows the results with 1-day, 2-days and 3-days of historical data, respectively. Similar to previous studies [14,38], we found that **time_of_day** (r = 0.11, 0.20, 0.16, and 0.14) and **day_of_week** (r = 0.06, 0.1, 0.08, and 0.08) had weak correlations with whether a participant was not drinking, drinking, or heavy drinking with no historical data, and with 1-, 2-, and 3-days of historical data, respectively. We found that these two temporal features (**time_of_day**, **day_of_week)**, and four mobile usage features (**screen_duration_interaction_seconds**, **average_time_between_keypress_ms**, **number_of_keypress_deletions**, and **number_of_keypress_insertions)** have positive relationships with drinking episodes with and without history data.

While all correlations were weak, we also had mixed results regarding the value of using historical data for correlating with different types of drinking episodes. For example, certain types of smart phone interactions (**happy_emoticon_count**, r = 0.14, 0.11, and 0.10) appeared in the top list of correlations for the analysis with 1-, 2-, 3-days of historical data, but not in the without historical data analysis. The movement features, **accelerometer_min_magnitude** (r = -0.05) and **accelerometer_mean_magnitude** (r = -0.03) and **accelerometer_median_magnitude** (r = -0.03) had a weak negative relationship with whether a participant was not drinking, drinking, or heavy drinking only when we *did not use* history data. In addition, **Radius_of_gyration** was correlated with drinking episodes when only using 3-day history.

We now provide more details on the correlation analysis, broken out by type of features: *e.g.*, temporal, motion, device usage and communications.

4.2 Day of Week and Time of Day

The correlation analysis (Table 5) shows that **time_of_day** and **day_of_week** correlate significantly with drinking episodes. As shown in Figure 3a and b, our results support prior research [68] showing that young adults tend to engage in heavy drinking from Thursday to Sunday, and were more likely to report drinking episodes on Saturday than other days, and during evening and night times compared to other times of the day. Heavy drinking episodes had longer average durations (M=8.96, SD=8.37) than drinking occasions (M=5.79, SD=5.21).



(a) Day of week (days) (b) Time of day (hours)

Fig. 3. **(a) Drinking episodes collected by ESM (Experience Sampling Method) via smartphones across all 30 users. (b) Time of day using 1-hour time window. 1 starts from 00:00 to 00:59 and 24 refers to the time between 23:00 to 23:59.**

4.3 Motion: activity and movement

The correlation analysis shows that there is a significant relationship between **mean**, **median** and **minimum magnitude of acceleration** and drinking episodes without historical data: r = -0.03, $p < .05$, r = -0.03, $p < .05$, and r = -0.05, $p < .001$ respectively and no relationship using historical data r = 0.02, $p = .3314$, r = 0.03, $p = .066$, and r = 0.03, $p = .0639$ respectively. This tells us that by using historical acceleration data, we are unlikely to be able to more easily differentiate non-drinking and drinking and heavy drinking episodes, compared to attempting this without using historical data.

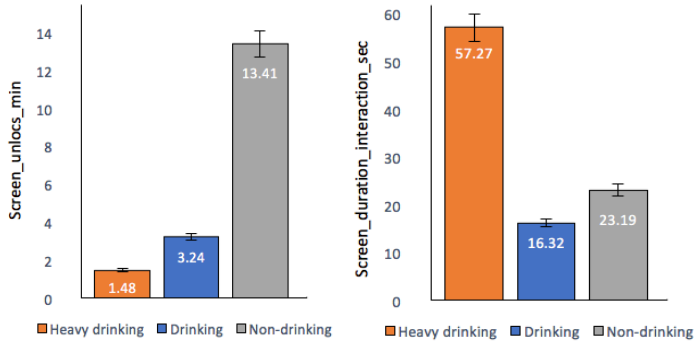
We also report on a few other features related to activity and movement. There is a significant positive relationship between **maximum magnitude of acceleration** and drinking episodes when not using historical data, r = 0.11, $p < .001$. The **number of activity changes**, meaning the number of transitions between different types of activities (still, tilting, walking, running, on bicycle, and in vehicle), correlates with drinking episodes (r = 0.03, $p < .05$), meaning that people who show an increase in physical activity changes are more likely to report a drinking event within a given time window. The **radius of gyration** meaning the radius of the circle which encompasses all of the places an individual visited in a given time segment [25] is also a good predictor to differentiate between non-drinking and drinking episodes both when not using historical data, r = 0, $p = .7799$, and using historical data r = 0.03, $p = .0535$. However, **travel distance meter** meaning the distance traveled in each time window, does not correlate with non-drinking and drinking episodes (r = 0, $p=.8263$).

4.4 Device usage: screen, unlock and keyboard operation

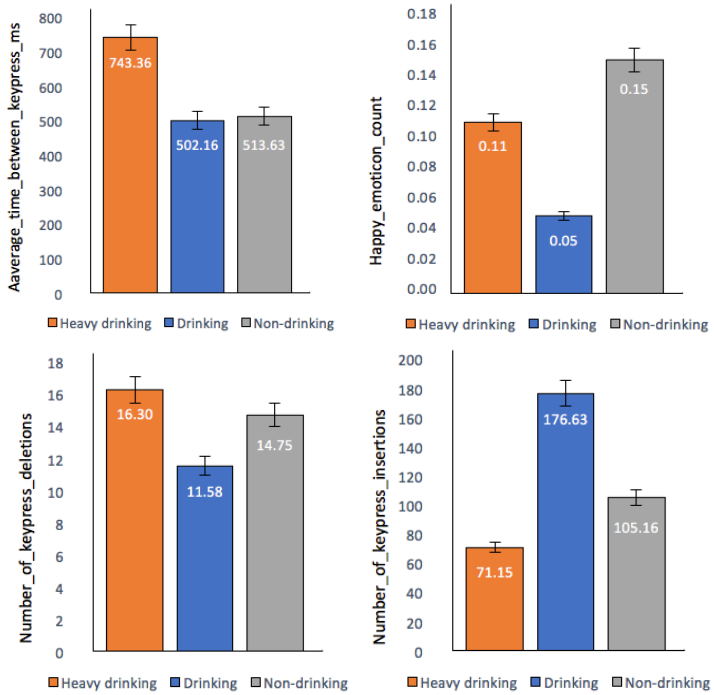
The results show that there is a significant relationship between **screen_duration_interaction_sec** (total duration of screen on and off within time segment), and drinking episodes when we use the previous 3 days of historical data before each drinking episode, r= 0.07, $p<0.001$. We found that there is a significant relationship without this historical data as well, r= 0.08, $p<0.001$. The **average time duration for interacting with screen** of the phone was lower when participants were drinking (M=16.32, SD=48.13) compared to non-drinking episodes (M=23.19, SD=77.13). However, the interaction durations of screen were higher for heavy drinking episodes (M=57.27, SD=127.31) when compared to non-drinking (M=23.19, SD=77.13) and drinking episodes (M=16.32, SD=48.13) (Figure 4a).

Interestingly, **screen_unlocks_per_minute** was likely to be lower when a participant was drinking (M=3.24, SD=27.37) and heavy drinking (M=1.48, SD=9.88) compared to non-drinking episodes (M=13.41, SD=467.40) (Figure 3a), which means people tend to check their smartphones less frequently during drinking events. This number was a little lower when participants were heavy drinking compared to when they were drinking (but still less than when they were not drinking).

Our results show that when participants were heavy drinking, they had longer average times between keyboard presses and a lower number of keypress insertions compared to when they were drinking or non-drinking (Figure 4b).



(a) Screen_unlocks_min (left) and duration_interaction_sec (right) between heavy drinking, drinking and non-drinking events



(b) Keypress interaction: average time between keypresses, happy emoticon count, number of keypress deletions and insertions

Fig. 4. **Users' device usage; screen, lock and keyboard between episodes; heavy drinking, drinking and non-drinking.**

4.5 Communication: calls and messages

Our results show that there is a positive relationship between the **number of correspondents** (*i.e.*, number of individuals with whom the participant communicated) and non-drinking and drinking episodes, $r = 0.03$, $p < .05$. However, the **number of incoming** and **outgoing messages** did not correlate with any drinking and non-drinking episodes ($r = 0.02$, $p = 0.0704$) and ($r = 0.02$, $p = 0.1475$) respectively.

Despite the number of significant correlations between these different types of features, all correlations were quite weak with the drinking episodes of interest. Further, our correlation analysis revealed somewhat conflicting evidence about the value of using historical data. While most of the feature correlations increased in magnitude with more historical data, adding more historical data was not uniformly positive.

Due to this conflicting evidence and the weak correlation, to find optimal features with which to build our models, we performed an Information Gain analysis on the feature set. We use Information Gain to select a smaller number of features than the maximum, to avoid overfitting to the data. In the analysis below, we target the features with the 20 highest information gain scores.

4.6 Feature Understanding: Information Gain

We measured the Information Gain (IG) of each sensor stream to understand the relationship between drinking episodes and the importance of behavioral sensor streams for detecting drinking episodes. Here we measure the Information Gain of each feature for each of our target classes; non-drinking, drinking and heavy drinking. From the IG, we chose the top ranked 20 features (IG > 0.05) across all our data. Figure 5 shows the results of our Information Gain analysis for a Random Forest classifier without (1h_no) and with historical data (1h_1, _2 and _3day) using a 1-hour data window size.

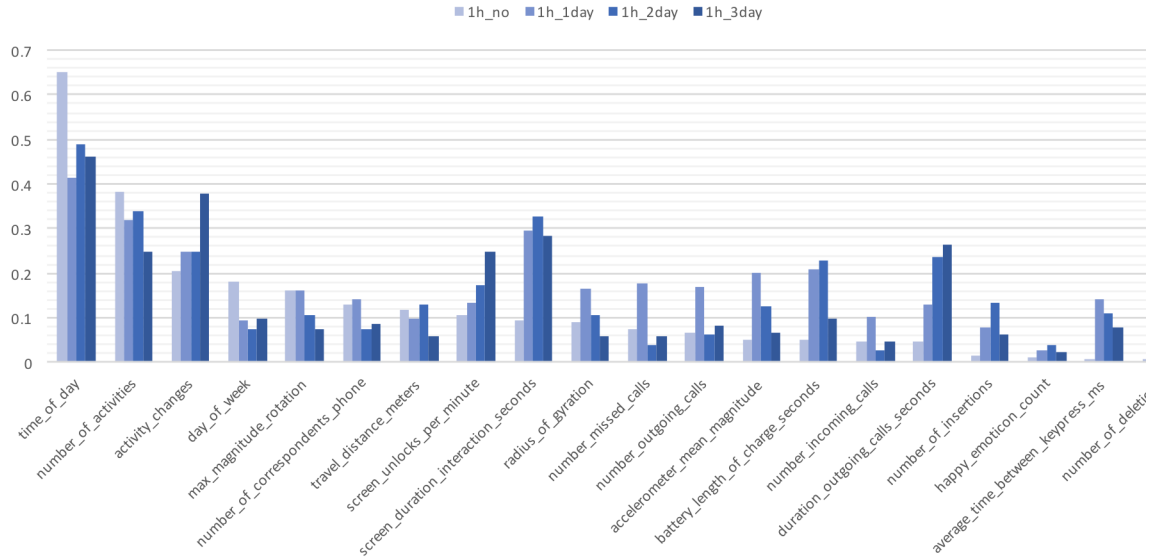


Fig. 5. **Information Gain (IG) of sensor streams with different amounts of historical data, using a 1-hour data window. Top ranked 20 features were used to build our non-, drinking and heavy drinking detection models.**

First, using these models, we confirmed that a traditional predictor of drinking behavior, time of day attribute is top ranked as 0.650, 0.413, 0.488 and 0.460, for the 1h_no, 1h_1day, 1h_2day and 1h_3day model respectively, across all users. Second, from our initial list of 56 features, movement features such as **number_of_activities** (0.383, 0.320, 0.339, and 0.249), activity changes (0.203, 0.248, 0.247, and 0.378) were quite good for detecting non-drinking, drinking and heavy drinking across all models. Third, the importance of smartphone usage features such as **screen_duration_interaction_seconds** increased (0.294, 0.327 and 0.282) with 1-, 2-, and 3-days of historical data compared to models without history data (0.096) to detect non-drinking, drinking and heavy drinking.

In addition, the **screen_unlocks_per_minute** and **battery_length_of_charge_seconds** features had information gain scores of 0.134, 0.173 and 0.247 and 0.210, 0.227 and 0.096, respectively, if historical data were added compared to when no-history data was used (0.104 and 0.050 respectively). Fourth, communication features, such as the **number of incoming calls**, **missed calls** (except for **number_of_correspondents_phone**) were relatively poor for identifying non-drinking, drinking and heavy drinking episodes, from the Information Gain perspective. However, the importance of communication features *e.g.*, **number_missed_calls** and **number_outgoing_calls** had the highest information gain scores (0.1765 and 0.1689, respectively) with 1-day of historical data. Fifth, when using a 2-hour window of historical data from smartphones, we found that more detailed keyboard measures such as **number_of_keypress_insertions** (0.134) and **number_of_keypress_deletions** (0.120) had relatively higher information gain scores than when not using historical data. Lastly, the importance of movement and location features **accelerometer_mean_magnitude** (0.200), **max_magnitude_rotation** (0.162) and **radius_of_gyration** (0.163) respectively had higher information gains with 1-day of historical data, compared to using no-history data. This means that movement and location sensors on smartphones, particularly with 1 day of historical day, can also contribute to accurately detecting drinking episodes.

Table 6. **Description of Information Gain (IG) Ranking Filter: Importance of Top 20 Ranked Attributes**

Ranking	Category	Features	Importance
1	Time	time_of_day	0.50338
2	Motion	number_of_activities	0.32338
3	Motion	count_changes_of_activities	0.26943
4	Device usage	screen_duration_interaction_seconds	0.25022
5	Communication	duration_outgoing_calls_seconds	0.16987
6	Device usage	screen_unlocks_per_minute	0.16493
7	Device usage	battery_length_of_charge_seconds	0.14618
8	Motion	max_magnitude_rotation	0.12541
9	Time	day_of_week	0.11206
10	Motion	accelerometer_mean_magnitude	0.11069
11	Communication	number_of_correspondents_phone	0.10837
12	Location	radius_of_gyration	0.10543
13	Location	travel_distance_meters	0.10008
14	Communication	number_outgoing_calls	0.09475
15	Communication	number_missed_calls	0.08767
16	Device usage	average_time_between_keypress_ms	0.08535
17	Device usage	number_of_deletions	0.08528
18	Device usage	number_of_insertions	0.07263
19	Communication	number_incoming_calls	0.05535
20	Device usage	happy_emoticon_count	0.02490

4.7 Classifier Performance

In this section, our goal is to optimize our drinking detection models, and to see the effects of window size and days of history on the model accuracies. We trained three machine-learning classifiers, Random Forest (RF), C4.5 decision tree and Bayesian network (BN) using the top ranked 20 features from our Information Gain analysis (Table 6). We used the F-score to optimize the models. Table 7 shows a detailed view of our classification comparisons having different data window sizes with averaged accuracies in classifying non-drinking, drinking and heavy drinking episodes using these metrics: Kappa, accuracy, precision, recall, F-score, MCC (Mathews Correlation Coefficient) and ROC (Receiver operating characteristics). Kappa is a measure of the similarity between observations and predictions while correcting for agreement which happens by chance [25]. ROC refers to a relation between the sensitivity (true positives) and the specificity (true negatives) of the algorithm [47]. MCC is a Pearson product-moment correlation coefficient between the observed and predicted classifications that can be used when dealing with unbalanced classes [30]. In addition, we performed two additional analyses to improve the quality of our models. First, we analyzed the impact of time window size on the model accuracy. The time window size refers to how recent the data is for calculating our features to determine optimal time windows for intervention delivery. If a 1-hour time window size is used, then a 1-hour snapshot of the smartphone sensor data is used to calculate the identified features. We experimented with 30-minute, 1-hour and 2-hour time windows to compare the accuracy of our model. Second, we analyzed the impact of historical data on the accuracy of our models to see how much data needs to be stored on the phone for accurate drinking detection. We use up to 3 days' worth of sensor data before the drinking events, to assess a model's accuracy.

Table 7. **Model Evaluation for Different Window Sizes and Number of Days of Historical Data: Random Forest (RF), C4.5 and Bayesian Network (BN). Metrics for each Model Include Averaged Kappa (K), Accuracy, Precision, Recall and F-score in Classifying. Bold indicates the top performing results for a particular metric.**

Model	Kappa	Accuracy	Precision	Recall	F-Score	MCC	ROC Area
30m_n_RF	0.339	0.946	0.936	0.946	0.94	0.377	0.871
30m_n_C4.5	0.336	0.914	0.938	0.914	0.925	0.381	0.818
30m_n_BN	0.043	0.928	0.904	0.929	0.916	0.045	0.729
30m_1d_RF	**0.842**	0.952	0.951	0.952	0.951	**0.838**	**0.976**
30m_1d_C4.5	0.719	0.904	0.921	0.904	0.91	0.723	0.895
30m_1d_BN	0.465	0.816	0.837	0.816	0.825	0.469	0.866
30m_2d_RF	0.757	0.948	0.948	0.949	0.948	0.752	0.967
30m_2d_C4.5	0.366	0.808	0.874	0.808	0.833	0.398	0.899
30m_2d_BN	0.297	0.9	0.905	0.901	0.87	0.395	0.882

30m_3d_RF	0.804	**0.966**	**0.966**	**0.967**	**0.966**	0.809	0.961
30m_3d_C4.5	0.778	0.96	0.962	0.961	0.961	0.785	0.939
30m_3d_BN	0.61	0.938	0.935	0.938	0.936	0.612	0.93

(a) 30-minute window size

Model	Kappa	Accuracy	Precision	Recall	F-Score	MCC	ROC Area
1h_n_RF	0.158	0.931	0.909	0.932	0.919	0.174	0.825
1h_n_C4.5	0.175	0.902	0.91	0.903	0.906	0.196	0.671
1h_n_BN	0.021	0.934	0.895	0.934	0.913	0.039	0.694
1h_1d_RF	0.732	0.916	0.92	0.917	0.918	0.732	0.96
1h_1d_C4.5	0.536	0.85	0.865	0.851	0.857	0.543	0.846
1h_1d_BN	0.444	0.833	0.845	0.833	0.837	0.446	0.821
1h_2d_RF	0.687	0.94	0.936	0.941	0.936	0.696	0.936
1h_2d_C4.5	0.581	0.907	0.911	0.907	0.909	0.599	0.79
1h_2d_BN	0.327	0.879	0.861	0.88	0.868	0.333	0.805
1h_3d_RF	0.717	0.955	0.953	0.956	0.953	0.719	0.946
1h_3d_C4.5	0.519	0.912	0.919	0.913	0.916	0.511	0.802
1h_3d_BN	0.304	0.888	0.877	0.889	0.883	0.293	0.729

(b) 1-hour window size

Model	Kappa	Accuracy	Precision	Recall	F-Score	MCC	ROC Area
2h_n_RF	0.09	0.926	0.902	0.926	0.913	0.127	0.819
2h_n_C4.5	0.218	0.91	0.917	0.91	0.913	0.246	0.576
2h_n_BN	0.005	0.931	0.892	0.931	0.911	0.026	0.762
2h_1d_RF	0.505	0.841	0.846	0.842	0.844	0.489	0.853
2h_1d_C4.5	0.438	0.823	0.824	0.824	0.823	0.43	0.758
2h_1d_BN	0.317	0.744	0.797	0.744	0.766	0.363	0.834
2h_2d_RF	0.588	0.922	0.915	0.923	0.916	0.587	0.94
2h_2d_C4.5	0.514	0.908	0.897	0.909	0.901	0.509	0.783
2h_2d_BN	0.453	0.87	0.888	0.87	0.878	0.467	0.769
2h_3d_RF	0.558	0.938	0.928	0.938	0.929	0.568	0.873
2h_3d_C4.5	0.415	0.903	0.9	0.903	0.902	0.404	0.751
2h_3d_BN	0.22	0.881	0.876	0.882	0.877	0.212	0.639

(c) 2-hour window size

Overall, the Random Forest (RF) model generally outperformed the C4.5 decision tree model, which generally outperformed the Bayesian network (BN) model, in terms of accuracy, for different window sizes of data and differing numbers of days of history (Figure 6). Our most accurate population-based classifier had an average classification accuracy of 96.6% across the data from all 30 young adults (average Kappa, accuracy, precision, recall, F-score, MCC and ROC: 0.804, 0.966, 0.966, 0.967, and 0.966, 0.809 and 0.961 respectively, Figure 7) in distinguishing non-drinking, drinking, and heavy drinking episodes. This model, *30m_3d_RF* (features calculated using 30-minute window sizes, with 3 days of historical data) outperformed the models with 1-hour (95.5%) and 2-hour (93.8%) windows and 3 days of historical data. However, we found that the *30m_1d_RF* model (30-minute window size for computing behavioral features with 1-day historical data) resulted in the highest kappa (k=0.842), MCC value (0.838) and ROC value (0.976) compared to the highest performance model (*30m_3d_RF* k=0.804, MCC: 0.809, ROC: 0.961), but slightly lower accuracy (95.2%) and F-score (0.951).

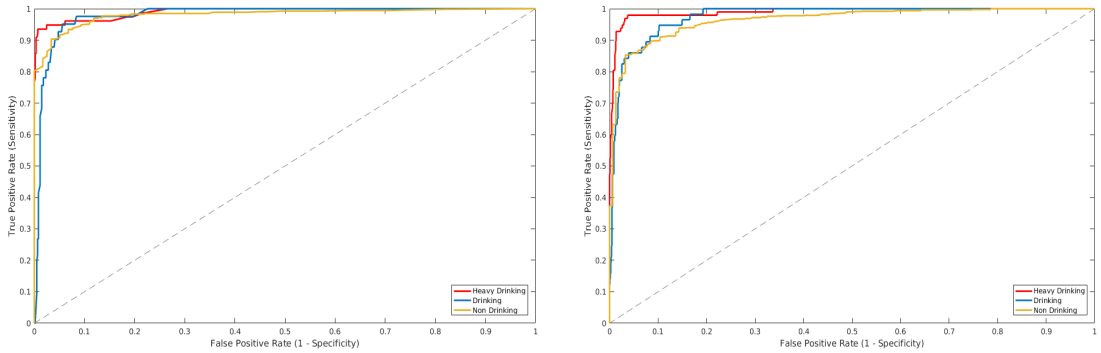
(a) 30-minute window size

(b) 1-hour window size

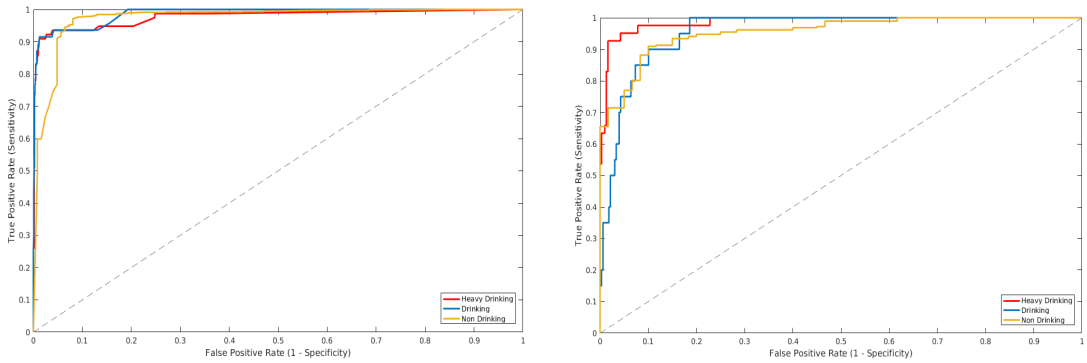
(c) 2-hour window size

Fig. 6. **Model performance comparisons using Random Forest (RF), C4.5 and Bayesian Network (BN) classifiers with 1-, 2 days and without historical data.**

We computed the ROC convex hulls (see Figure 7) for the best performing models (ROC area >= 0.96) because the curves represent the subset of the best decision boundaries showing the relative costs of true positives (Y-axis represents *sensitivity*) and false positives (X-axis represents 1- *specificity*). (0, 1) on the ROC curve would be an ideal point. They indicate the relative tradeoffs that can be made when tuning for a particular true positive and false positive balance. As expected, the false positive rates are low for high true positive rates, particularly for the two models with the highest accuracy *30m_3d_RF* (Figure 7a) and *30m_3d_RF* (Figure 7c). In addition, for detecting heavy drinking, the area under the curve is highest (0.992) for the *30m_1d_RF* model, and for detecting drinking, it is highest (0.948) for the *30m_3d_RF* model.



(a) 30-minute window size – 1day historical data (30m_1d_RF) (b) 30min – 2days historical data (30m_2d_RF)



(c) 30-minute window size – 3days historical data (30m_3d_RF) (d) 1hour window size – 1day historical data (1h_1d_RF)

Fig. 7. **Receiver Operating Characteristics(ROC) curves for the best performing models with Random Forest(RF)**

4.8 Classifier Performance for Different Window Sizes

As Figure 8 shows, Kappa decreases when the data window size increases. Correspondingly, the false positive rate increases, and the true positive rate slightly decreases as the data window size increases. The accuracy of each model (meaning 30min, 1h-, and 2h-window) trained and tested with no, 1-day, 2-day and 3-day historical data all have similar trends (Figure 9a, b, c). The results show that the optimal window size for detecting non-drinking, drinking and heavy drinking was when the social and behavioral data were collected within a 30-minute window.

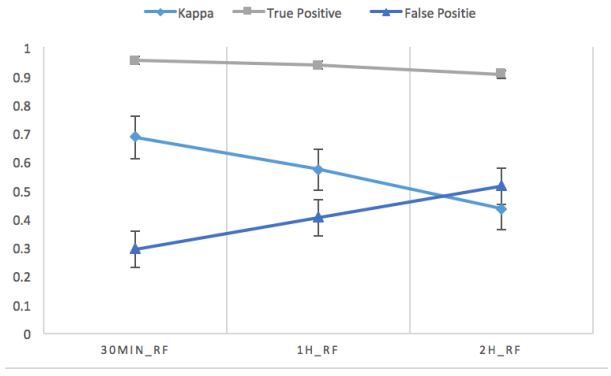
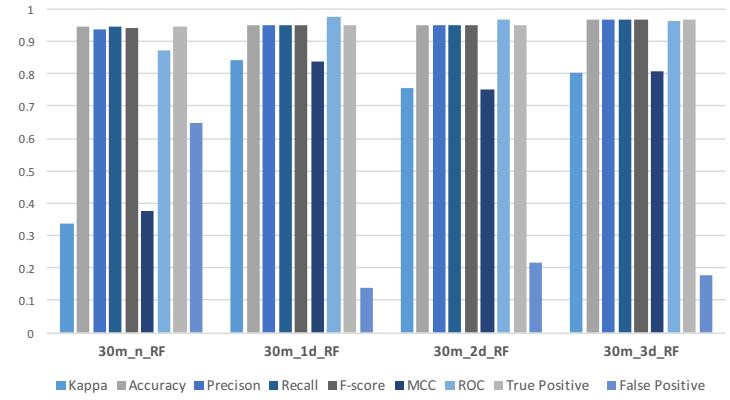
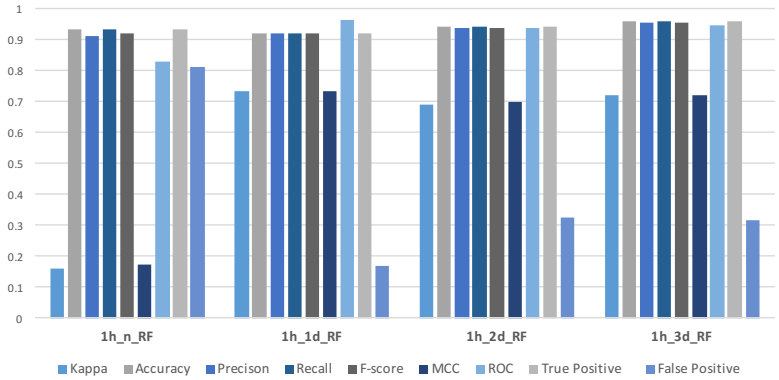


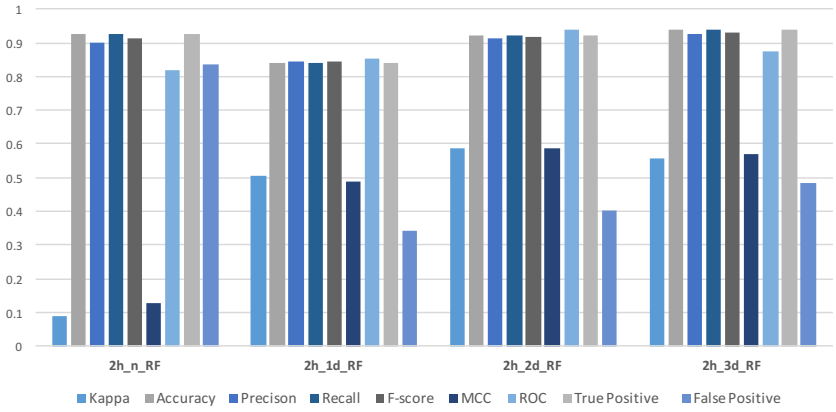
Fig. 8. **RF Model comparisons of Kappa, True Positive, and False positive rates: An average of each window with historical data; no-days, 1-day, 2-days, and 3-days data.**



(a) 30-minute window size



(b) 1-hour window size



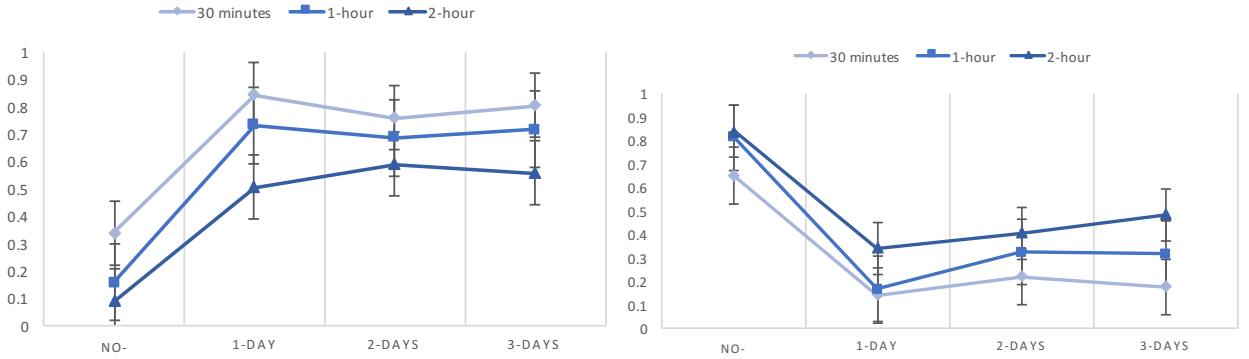
(c) 2-hour window size

Fig. 9. The results of the model comparisons differentiated window sizes as well as with and without historical data in which we used the Random Forest (RF) classifier to detect drinking episodes because the RF showed the best performance among models in the previous experiments. In specific, model comparisons in classifying 30 minutes (30m), 1-hour (1h) and 2-hour (2h) windows as

non-drinking (N), heavy drinking (H) and drinking (D), and drinking detection results without (n)- and with 1day (1d), 2days (2d) and 3days (3d) of historical data. The metrics used are Kappa, accuracy, precision, recall, F-score, MCC, ROC and true positive (TP) and false positive (FP).

4.9 Classifier Performance for Different Amounts of Historical Data

We now focus on the results of the models built with differing amounts of historical data: no-, 1-, 2- and 3-day histories. The highest *Kappa* value, 0.842, was achieved with the *30m_1d_RF* model. Landis and Koch [33] define values in the range of 0.81-1 as almost perfect, and Fleiss [25] defines values > 0.75 as excellent. Using historical data results in higher values of *Kappa*, as shown in Figure 10a. It also results in lower false positive rate compared to models trained without using historical data of the smartphones (Figure 10b), regardless of the data window size chosen. Further, accuracy of models with 3-days historical data, when compared to models with no historical data, was higher for all data window sizes (30m: 96.6% vs. 94.6%; 1h: 95.5% vs. 93.1%; 2h: 93.8% vs. 92.6%).



(a) **Kappa** (b) **False Positive (FP)**

Fig. 10. **Comparison of the RF models trained on different amounts of historical data (no-days, 1-day, 2-days, and 3-days), averaged across the different data window sizes.**

4.10 Classifier Performance for Heavy Drinking

When it comes to delivering an appropriate intervention, we would like to be able to predict heavy drinking episodes, and not just detect them. However, detection of these heavy drinking episodes is a necessary stepping stone and is valuable in and of itself. As described when we motivated our work in the introduction, when a heavy drinking episode is detected for an individual, an automated system can, for example, send an intervention message to alert designated individuals to assist or watch out for this person, or send a *post-hoc* message providing feedback to the individual regarding recent heavy drinking episodes. The latter supports both self-reflection and possible changes to a care plan by a clinician (if the individual is in treatment). To enable such features, we need to **maximize the true positive rate for detecting heavy drinking episodes (**as opposed to maximizing the accuracy or true positive rate across all 3 types of episodes). The consequences of an incorrect prediction (in particular, a false negative) could involve, for example, a missed opportunity to provide real-time support to reduce alcohol-related harm, whereas a false positive in the context of a mobile intervention (*e.g.*, sending a message that encourages a halt to drinking) could erode engagement with the intervention [57].

Figure 11 shows the True Positive Rate (TPR) and False Positive Rate (FPR) for classifying all three classes (non-drinking, drinking, heavy drinking). Our results show that models using 1 day of historical data had the highest true positive rate for classifying heavy drinking. Among these, the *30m_1d_RF* model represents the highest TPR for classifying heavy drinking (0.909). As expected, all models using historical data performed better than with no historical data.

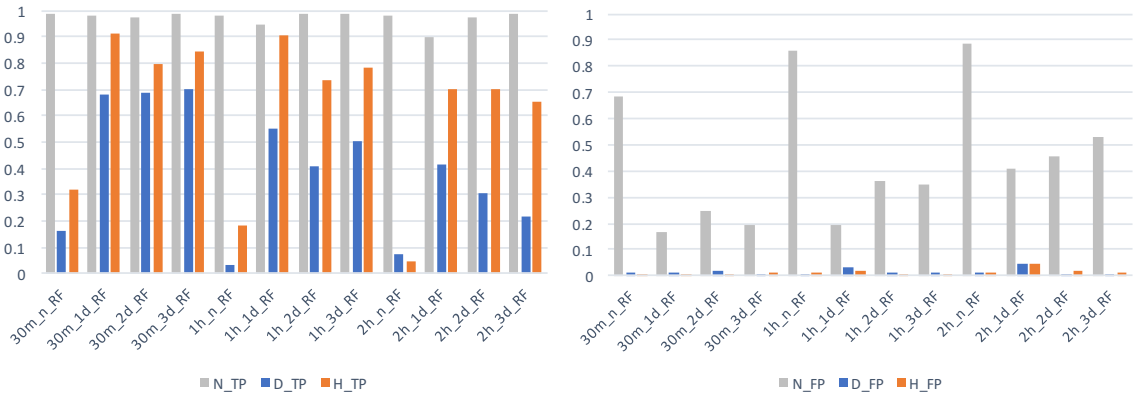


Fig. 11. **True Positive(TP) and False Positive(FP) in classifying non-drinking (N), drinking (D)and heavy drinking (H) episode.**

Table 8. **Metrics for each Model Include True Positive (TP) and False Positive (FP) in Classification: Model Evaluation for Different Window Sizes (30, 1h, 2h) and with (1d, 2d, 3d, where d is days) and without historical data (n) using Random Forest (RF).**

	30m_n_RF	30m_1d_RF	30m_2d_RF	30m_3d_RF	1h_n_RF	1h_1d_RF	1h_2d_RF	1h_3d_RF	2h_n_RF	2h_1d_RF	2h_2d_RF	2h_3d_RF
N_TP	0.984	0.979	0.975	0.985	0.980	0.944	0.983	**0.987**	0.978	0.897	0.972	0.985
D_TP	0.163	0.683	0.684	**0.702**	0.036	0.550	0.409	0.500	0.071	0.417	0.308	0.214
H_TP	0.321	**0.909**	0.794	0.844	0.186	0.902	0.732	0.783	0.048	0.700	0.700	0.650
N_FP	0.685	**0.169**	0.247	0.194	0.859	**0.197**	0.365	0.348	0.886	0.406	0.455	0.529
D_FP	0.010	0.013	0.016	0.006	0.005	0.034	0.012	0.009	0.013	0.048	0.007	0.006
H_FP	0.008	**0.005**	0.007	0.010	0.015	0.020	0.006	0.003	0.012	0.045	0.019	0.009

We also found that models using 1 day of historical data had the lowest false positive rate (false alarm) in detecting non-drinking. False positives for detecting non-drinking may have the highest cost to participants as they result in missed opportunities for interventions. Among these, the *30m_1d_RF* model represents the lowest FPR for classifying non-drinking (0.169) episodes among all our RF models.

4.10.1 Confusion Matrix

As Table 9b shows, the *30min_1d_RF* model performs the best at classifying heavy drinking: 90.9%. Its performance in classifying all episodes (non-drinking, drinking and heavy drinking) is 95.2%. It is interesting to note that classifiers incorrectly labelled heavy drinking episodes as non-drinking episodes for most of the mislabeled cases. With more historical data, the classifiers more correctly label heavy drinking episodes as heavy drinking. For example, with just 1 day of historical data (Table 9b), the true positive rate for heavy drinking rises from 32.1% (no history, Table 9a) to 90.9%. The same is true for the true positive rates for drinking episodes – more history improves accuracy.

Table 9. **Confusion Matrix per Model (Random Forest)**

		Estimated			
		Non-drinking	Drinking	Heavy drinking	Total
Actual	Non-drinking	**2322 (98.4%)**	22	16	2360
	Drinking	37	**8 (16.3%)**	4	49
	Heavy drinking	52	3	**26 (32.1%)**	81
	Total	2411	33	46	2490

(a) 30-minute window size, without historical data

		Estimated			
		Non-drinking	Drinking	Heavy drinking	Total
Actual	Non-drinking	**519 (98.4%)**	8	3	530
	Drinking	13	**28 (68.3%)**	0	41
	Heavy drinking	7	0	**70 (90.9%)**	77
	Total	539	36	73	648

(b) 30-minute window size – 1day historical data used

		Estimated			
		Non-drinking	Drinking	Heavy drinking	Total
Actual	Non-drinking	**1100 (97.5%)**	20	8	1128
	Drinking	18	**39 (68.4%)**	0	57
	Heavy drinking	20	0	**77 (79.4%)**	97
	Total	1138	59	85	1282

(c) 30-minute window size – 2days historical data used

		Estimated			
		Non-drinking	Drinking	Heavy drinking	Total
Actual	Non-drinking	**1191 (98.5%)**	7	11	1209
	Drinking	13	**33 (70.2%)**	1	47
	Heavy drinking	11	1	**65 (84.4%)**	77
	Total	1215	41	77	1333

(d) 30-minute window size – 3days historical data used

		Estimated			
		Non-drinking	Drinking	Heavy drinking	Total
Actual	Non-drinking	**1151 (98%)**	6	17	1174
	Drinking	26	**1 (3.6%)**	1	28

Heavy drinking	35	0	**8 (18.6%)**	43
Total	1212	7	26	1245

(e) 1-hour window size – no historical data used

		Estimated			
		Non-drinking	Drinking	Heavy drinking	Total
Actual	Non-drinking	**271 (94.4%)**	11	5	287
	Drinking	8	**11 (55%)**	1	20
	Heavy drinking	4	0	**37 (90.2%)**	41
	Total	283	22	43	348

(f) 1-hour window size – 1day historical data used

		Estimated			
		Non-drinking	Drinking	Heavy drinking	Total
Actual	Non-drinking	**469 (98.3%)**	6	2	477
	Drinking	12	**9 (40.9%)**	1	22
	Heavy drinking	11	0	**30 (73.2%)**	41
	Total	492	15	33	540

(g) 1-hour window size – 2days historical data used

		Estimated			
		Non-drinking	Drinking	Heavy drinking	Total
Actual	Non-drinking	**627 (98.7%)**	6	2	635
	Drinking	12	**12 (50%)**	0	24
	Heavy drinking	11	0	**31 (78.3%)**	42
	Total	650	18	33	701

(h) 1-hour window size – 3days historical data used

		Estimated			
		Non-drinking	Drinking	Heavy drinking	Total
Actual	Non-drinking	**567 (97.8%)**	7	6	580
	Drinking	12	**1 (7.1%)**	1	14
	Heavy drinking	19	1	**1 (4.8%)**	21
	Total	598	9	8	615

(i) 2-hour window size – without historical data

		Estimated			
		Non-drinking	Drinking	Heavy drinking	Total
Actual	Non-drinking	**130 (89.7%)**	8	7	145
	Drinking	7	**5 (41.7%)**	0	12
	Heavy drinking	6	0	**14 (70%)**	20
	Total	143	13	21	177

(j) 2-hour window size – 1day historical data used

		Estimated			
		Non-drinking	Drinking	Heavy drinking	Total
Actual	Non-drinking	**245 (97.2%)**	2	5	252
	Drinking	9	**4 (30.8%)**	0	13
	Heavy drinking	6	0	**14 (70%)**	20
	Total	260	6	19	285

(k) 2-hour window size – 2days historical data used

		Estimated			
		Non-drinking	Drinking	Heavy drinking	Total
Actual	Non-drinking	**333 (98.5%)**	2	3	338
	Drinking	11	**3 (21.4%)**	0	14
	Heavy drinking	7	0	**13 (65%)**	20
	Total	351	5	16	372

(l) 2-hour window size – 3days historical data used

5 DISCUSSION

Our results show that the use of only passively collected data from smartphones can capture young adults' drinking and particularly their heavy drinking behaviors. Smartphones, which are almost always owned and carried by this population, provide an opportunity to detect drinking and heavy drinking behaviors using data streams such as physical movements, location, phone usage and communications without having to rely on the burdensome self-report of number of drinks and durations of drinking episodes. Importantly, our drinking detection model was developed primarily to support young adults with strategies during drinking episodes (to limit drinking and to provide help) and opportunities to reflect on drinking episodes and patterns after the fact. We optimized the model to detect drinking and heavy drinking episodes, in terms of the data features used, the window size for performing data analysis, and the number of days of historical data. The model was not developed to detect alcohol use for other purposes (*e.g.*, forensic or legal), since these purposes may involve minimizing different types of errors.

Our work examined the utility of smartphone sensors to track certain physical and social behaviors that are associated with drinking episodes in young adults. Smartphone sensors can capture physical activity, but they also can record social communication activity, such as phone usage (*e.g.*, calls, messages) and screen on/off. In particular, screen on/off could be an important marker of heavy drinking episodes, with implications for signaling optimal times for intervention in future work. Our study provides insights about how much data needs to be collected on smartphones to increase the accuracy of heavy drinking and drinking detection (30 minute windows, with 3 days of historical data for overall classification and 30 minute windows with 1 day of classification for heavy drinking detection). We created a behavioral model that distinguishes non-drinking, drinking and heavy drinking based on both the physical behavior and social behavior collected using smartphone sensor data, with an accuracy of 96.6%. Our best model for detecting heavy drinking episodes had an accuracy of 95.2%

Our detection model outperforms existing drinking detection models that also use commodity hardware. For example, a lab study using the accelerometer on a smartphone to estimate gait for six participants had an accuracy of roughly 70% in detecting level of alcohol use (0-2, 3-6 or more than 6 drinks) [2]. A similar alcohol detection model that tracked data from smartwatch sensors including the accelerometer, gyroscope, heart-rate and skin temperature, had a precision of .886 for the binary classification problem of blood alcohol level above or below 0.0685 [27]. Recall that the precision of our best model was 0.966. Although these lab studies demonstrate the use of sensors to detect alcohol use, our work provides important advances in terms of larger sample size, detection of drinking and heavy drinking episodes "in the wild" over 28 days, rather than a single lab visit, and the novel use of smartphone sensors to detect social (*e.g.*, communication activities) and behavioral (*e.g.*, user-device interactions, such as key strokes; travel activity) markers associated with drinking episodes. Of note, our study shows that historical data, which were not available in the lab studies, improved the performance of detecting drinking and heavy drinking episodes in the natural environment.

With an accurate detection method, we can now begin to explore the use of intervention strategies that depend on detecting drinking and heavy drinking episodes. In particular, in the moment when drinking is detected, protective behavioral strategies can be delivered to slow down or halt the rate of drinking, particularly for those that are prone to transition from drinking to heavy drinking. If heavy drinking is detected, messages could similarly be delivered to the drinker, or to designated individuals to provide support for the drinker. After heavy drinking episodes, we can use visualization techniques to help young adults better reflect on their drinking patterns and use motivational strategies to encourage regulation of drinking patterns. In addition, clinicians could use recent drinking patterns to alter care plans for at-risk drinkers. The drinking models we have presented in this paper will allow us to test the efficacy of these approaches for reducing the incidence and cost of drinking. In the future, we want to build upon our detection work and try to predict future events of drinking and heavy drinking. This would allow us to proactively engage with individuals before they begin drinking or heavy drinking, through the user of just-in-time (or optimally timed) interventions.

For detection and prediction, it is important to reduce the false positive and false negative rates as much as possible. As an extreme example, the false positive rate of over 90% (for the *1h_n_BN* model) is too high for many applications. Fortunately, the false positive rates for drinking detection were much lower for our best models. False positives (in which the model identifies drinking when no drinking has occurred) may be preferred to false negatives (not identifying drinking events that actually occurred). Specifically, if the model detects drinking when none occurred, and a message encouraging strategies to reduce alcohol-related harm is sent (*e.g.*, "alternate alcoholic drinks with water"), the respondent may, for example, not understand why the message is being sent, since the message appears to be "out of context". Alternatively, if the model fails to detect drinking, and no intervention message is sent, an opportunity to address alcohol use proximal to its occurrence has been lost (which is a more serious error in the specific context of an intervention). For other applications, such as those involving detection of alcohol use for legal or court purposes, false positives could result in imposing erroneous sanctions.

More work is needed to identify other beneficial raw or computed values from smartphone sensor data, since only 56 sensor features were explored here. In addition, other indicators of the social and physical environment such as scheduled events, social media posts and social ties that are related to situations when young adults drink could be used to understand individual interests and social interactions, and to increase detection of drinking episodes, accordingly.

5.1 Real-time Intervention to Prevent Negative Consequences in a Timely Manner

Our development of the drinking detection model using only smartphone sensors is a first step toward "just in time" intervention. That is, first, the outcome of interest needs to be "detected" with reasonable accuracy. The next step will be to see if we can go beyond *detecting* drinking episodes, to *predicting* drinking and heavy drinking episodes. If these episodes can be predicted accurately, then intervention messages can be sent to individuals at a time when they appear to be "at risk" for heavy drinking. For example, prior to a drinking episode, a message encouraging non-drinking activities could be sent, whereas during a drinking episode, a message that encourages moderate drinking (*e.g.*, let a couple of hours pass before your next drink) could be delivered. Our ultimate goal is to use data collected by the smartphone, with minimal participant burden, to predict drinking and heavy drinking episodes for the purpose of informing "just in time" intervention delivery.

As further evidence of the value of appropriately timing messages, a recent proof-of-concept study [56], which experimentally manipulated the timing of mobile reminders to use stress reduction strategies, found that a group receiving proximal (*i.e.*, delivered in response to a self-report of high stress, or after detection) intervention messages had better outcomes (*e.g.*, lower self-reported stress, lower salivary cortisol) than a group receiving interventions at random times. This study demonstrates that proximal delivery of intervention messages is associated with better outcomes, emphasizing the utility of determining optimal timing for message delivery to improve intervention effects. Our future work will leverage message delivery, in which messages are timed to be sent prior to (in the predictive case), and during (in the detection case), drinking episodes to improve the effects of our alcohol intervention messaging.

5.2 Maximize Efficiency of Using Embedded Sensors on Smartphones for Detecting Drinking Episodes

While our application for collecting sensor data was not overly power-intensive (at least, none of our subjects complained about battery life), it could be made more efficient. We would both like to explore the use of additional sensor streams, while reducing the sampling rate of all the used sensor streams to reduce battery usage. This will be more important when our models for detection (and in the future, prediction) need to run on the phone, as the model execution will consume additional battery power. From a data storage perspective, as we obtained our strongest results with 1 or 3 days of history data, we can safely have our application always delete data older than 3 days. This policy also has positive privacy implications.

6 LIMITATIONS

Our work, although avoiding the limitations of other smartphone and wearable device studies, had some limitations of its own. First, while we used a larger group of participants than past work (n=30), it is still relatively small, so our model might have limited generalizability. Second, young adult participants showed reduced compliance toward the end of the 28-day data collection period. Compensation ($2 per completed report) was provided, but this micropayment schedule could be improved, for example, by using bonuses or other incentives to motivate consistent completion to improve ground truth data collection. Although self-reports of alcohol use collected using ESM is state-of-the-science and has demonstrated validity, self-reports are subject to possible bias (*e.g.*, under- or over-reporting). Perhaps combining self-reports with a transdermal alcohol monitor would allow us to capture the benefits of both for obtaining better ground truth data. Third, our application for heavy drinking detection allowed users to disable sensors (*e.g.*, Wi-Fi, Bluetooth, location) if they felt that battery drain was too high, or did not have enough storage space on their phones. A future deployable system would have to block users from disabling sensors, to not impact the system's ability to detect drinking episodes.

7 FUTURE WORK

Our immediate next work will be to study the impact of delivering different interventions based on drinking and heavy drinking detection. In addition, to increase the generalizability of our work, we plan to extend this line of research to other populations (*e.g.*, older individuals). The models we developed in the work were population models (using data from all participants). As a next step, we aim to build individual models for detecting drinking behavior as young adults are likely to exhibit different and mutable patterns over time when drinking. Our goal then, is to improve overall detection performance by leveraging individual differences. Finally, as stated previously, we would also like to expand our work to move beyond detection to prediction of heavy drinking, which would enable just-in-time intervention message delivery prior to initiation of a drinking episode, to increase message impact.

8 CONCLUSIONS

In this paper, we built a machine learning based-model that can detect whether an individual is not drinking, drinking, or heavy drinking with an accuracy of 96.6%, using smartphone data from participants. We identified the most important features for performing this detection, which can be used to deliver appropriate interventions either in the moment or after the drinking episodes, to reduce the frequency and severity of heavy drinking. We identify the relative value of using different amounts of historical data, and different size windows of data on detection accuracy, and the tradeoffs that can be made for balancing false positive and false negative rates. Our work provides guidance regarding how much data needs to be collected from smartphones to increase the accuracy of drinking and heavy drinking detection. Our results can be used to improve the timing of mobile intervention delivery, and is a first important step towards future work to predict drinking and heavy drinking episodes.

ACKNOWLEDGEMENTS

The authors acknowledge support by the National Institute of Alcohol Abuse and Alcoholism (NIAAA) under grants K23 AA023284-01 and R01 AA023650, and partial funding by the Academy of Finland (Grants 276786-AWARE, 286386-CPDSS, 285459-iSCIENCE, 304925-CARE), the European Commission (Grant, 6AIKA-A71143-AKAI), and Marie Skłodowska-Curie Actions (645706-GRAGE). We thank our UbiComp lab members and visiting researchers. We thank Afsaneh Doryab, Michael Merrill and Deepika Bablani for their help with data processing. We also thank Yuuki Nishiyama and Jung-Wook Park for their help with developments of the AWARE application, and Julian Ramos and Kyung-Joong Kim for discussions about machine learning.

REFERENCES

[1] A. Abbey. 2002. Alcohol-related sexual assault: A common problem among college students. Journal of Studies on Alcohol (14). 118-128.
[2] Zachary Arnold, Danielle Larose and Emmanuel Agu. Year. Smartphone inference of alcohol consumption levels from gait. In Healthcare Informatics (ICHI), 2015 International Conference on, IEEE, 417-426.
[3] Nikola Banovic, Christina Brant, Jennifer Mankoff and Anind Dey. 2014. ProactiveTasks: The Short of Mobile Device Use Sessions. In Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices &\#38; Services, ACM, 243-252. DOI: https://doi.org/10.1145/2628363.2628380
[4] Nancy P Barnett, EB Meade and Tiffany R Glynn. 2014. Predictors of detection of alcohol use episodes using a transdermal alcohol sensor. Experimental and clinical psychopharmacology, 22 (1). 86.
[5] Lauren Boggs, Jacquie Harris, Kristin Hays and Maggie Young. 2008. The Effects Of Binge Drinking Among College Students. Journal of Undergraduate Research.
[6] Linda Bolier and Wilhelmus Johannes Maria Josephus Cuijpers. 2000. Effectieve verslavingspreventie op school, in het gezin en in de wijk (Effective Drug Prevention at School, in the Family and in the Community). GGZ Nederland.
[7] Katharine A Bradley, Anna F DeBenedetti, Robert J Volk, Emily C Williams, Danielle Frank and Daniel R Kivlahan. 2007. AUDIT-C as a brief screen for alcohol misuse in primary care. Alcoholism: Clinical and Experimental Research, 31 (7). 1208-1217.
[8] Thomas K. Greenfield Canada, Bond Jason, C. Kerr William and Council Government of Canada. National Research. 2011. Biomonitoring for Improving Alcohol Consumption Surveys: The New Gold Standard?
[9] Kate B Carey and John TP Hustad. 2002. Are retrospectively reconstructed blood alcohol concentrations accurate? Preliminary results from a field study. Journal of Studies on Alcohol, 63 (6). 762-766.
[10] Stephanie Carreiro, Hua Fang, Jianying Zhang, Kelley Wittbold, Shicheng Weng, Rachel Mullins, David Smelson and Edward W Boyer. 2015. iMStrong: deployment of a biosensor system to detect cocaine use. Journal of medical systems, 39 (12). 186.
[11] CDC, CDC - Fact Sheets-Binge Drinking - Alcohol. Accessed from https://www.cdc.gov/alcohol/fact-sheets/binge-drinking.htm
[12] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16. 321-357.
[13] Gokul Chittaranjan, Jan Blom and Daniel Gatica-Perez. 2013. Mining large-scale smartphone data for personality studies. Personal and Ubiquitous Computing, 17 (3). 433-450.
[14] John D Clapp, James Lange, Jon Wong Min, Audrey Shillington, Mark Johnson and Robert Voas. 2003. Two studies examining environmental predictors of heavy drinking by college students. Prevention Science, 4 (2). 99-108.
[15] SF Cutler, PG Wallace and AP Haines. 1988. Assessing alcohol consumption in general practice patients—a comparison between questionnaire and interview (findings of the Medical Research Council's general practice research framework study on lifestyle and health). Alcohol and Alcoholism, 23 (6). 441-450.
[16] Jonathan D'Angelo, Bradley Kerr and Megan A Moreno. 2014. Facebook displays as predictors of binge drinking: From the virtual to the visceral. Bulletin of science, technology & society, 34 (5-6). 159-169.
[17] Lloyd D. Johnston, Patrick M. O'Malley, Richard A. Miech, Jerald G. Bachman, John E. Schulenberg. 2016. Monitoring the Future national survey results on drug use, 1975–2015: Volume 2, College students and adults ages 19–55. Ann Arbor: Institute for Social Research, The University of Michigan. Available at http://monitoringthefuture.org/pubs.html#monographs
[18] Trinh Minh Tri Do and Daniel Gatica-Perez. Year. Groupus: Smartphone proximity data and human interaction type mining. In Wearable Computers (ISWC), 2011 15th Annual International Symposium on, IEEE, 21-28.
[19] Carolin Donath, Elmar Gräßel, Dirk Baier, Christian Pfeiffer, Stefan Bleich and Thomas Hillemacher. 2012. Predictors of binge drinking in adolescents: ultimate and distal factors-a representative study. BMC public health, 12 (1). 263.
[20] David H Epstein, Matthew Tyburski, Ian M Craig, Karran A Phillips, Michelle L Jobes, Massoud Vahabzadeh, Mustapha Mezghanni, Jia-Ling Lin, C Debra M Furr-Holden and Kenzie L Preston. 2014. Real-time tracking of neighborhood surroundings and mood in urban drug misusers: application of a new method to study behavior in its geographical context. Drug and alcohol dependence, 134. 22-29.
[21] George Fein and David Greenstein. 2013. Gait and balance deficits in chronic alcoholics: No improvement from 10 weeks through 1 year abstinence. Alcoholism: Clinical and Experimental Research, 37 (1). 86-95.
[22] D. Ferreira, J. Goncalves, V. Kostakos, L. Barkhuus and A. K. Dey. 2014. Contextual Experience Sampling of Mobile Application Micro-Usage. In International Conference on Human-Computer Interaction with Mobile Devices and Services, 91-100. DOI: https://doi.org/10.1145/2628363.2628367
[23] D. Ferreira, V. Kostakos, A. R. Beresford, J. Lindqvist and A. K. Dey. 2015. Securacy: An Empirical Investigation of Android Applications' Network Usage, Privacy and Security. In Conference on Security and Privacy in Wireless and Mobile Networks, 11 11-11 11. DOI: https://doi.org/10.1145/2766498.2766506
[24] Denzil Ferreira, Vassilis Kostakos and Anind K. Dey. 2015. AWARE: Mobile Context Instrumentation Framework. Frontiers in ICT, 2 (6). 1-9. DOI: https://doi.org/10.3389/fict.2015.00006
[25] Joseph L Fleiss, Bruce Levin and Myunghee Cho Paik. 2013. Statistical methods for rates and proportions. John Wiley & Sons.
[26] Thomas K Greenfield, Madhabika B Nayak, Jason Bond, William C Kerr and Yu Ye. 2014. Test–retest reliability and validity of life-course alcohol consumption measures: the 2005 National Alcohol Survey follow-up. Alcoholism: clinical and experimental research, 38 (9). 2479-2487.
[27] Mario A Gutierrez, Michelle L Fast, Anne H Ngu and Byron J Gao. Year. Real-Time Prediction of Blood Alcohol Content Using Smartwatch Sensor Data. In International Conference on Smart Health, Springer, 175-186.
[28] Martin Hagger, G Wong and S Davey. 2015. A theory-based behavior-change intervention to reduce alcohol consumption in undergraduate students: Trial protocol Health behavior, health promotion and society.
[29] Nathalie Hill-Kapturczak, John D Roache, Yuanyuan Liang, Tara E Karns, Sharon E Cates and Donald M Dougherty. 2015. Accounting for sex-related differences in the estimation of breath alcohol concentrations using transdermal alcohol monitoring. Psychopharmacology, 232 (1). 115-123.
[30] Giuseppe Jurman, Samantha Riccadonna and Cesare Furlanello. 2012. A comparison of MCC and CEN error measures in multi-class prediction. PloS one, 7 (8). e41882.
[31] John T Kent. 1983. Information gain and a general measure of correlation. Biometrika, 70 (1). 163-173.
[32] SeungJun Kim, Jaemin Chun and Anind K Dey. Year. Sensors know when to interrupt you in the car: Detecting driver interruptibility through monitoring of peripheral interactions. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, 487-496.
[33] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. biometrics. 159-174.
[34] Thad R Leffingwell, Nathaniel J Cooney, James G Murphy, Susan Luczak, Gary Rosen, Donald M Dougherty and Nancy P Barnett. 2013. Continuous objective monitoring of alcohol use: twenty-first century measurement using transdermal sensors. Alcoholism: Clinical and Experimental Research, 37 (1). 16-22.
[35] P. Leroux, K. Roobroeck, B. Dhoedt, P. Demeester and F. De Turck. 2013. Mobile application usage prediction through context-based learning. Journal of Ambient Intelligence and Smart Environments, 5 (2). 213-235. DOI: https://doi.org/10.3233/Ais-130199
[36] Jun-Ki K. Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman and Jason I. Hong. 2014. Toss 'N' Turn: Smartphone As Sleep and Sleep Quality Detector. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 477-486. DOI: https://doi.org/10.1145/2556288.2557220

[37] Frederick Muench, Katherine van Stolk-Cooke, Alexis Kuerbis, Gertraud Stadler, Amit Baumel, Sijing Shao, James R McKay and Jon Morgenstern. 2017. A Randomized Controlled Pilot Trial of Different Mobile Messaging Interventions for Problem Drinking Compared to Weekly Drink Tracking. PloS one, 12 (2). e0167900.
[38] Mark Muraven, R. Lorraine Collins, Saul Shiffman and Jean A. Paty. 2005. Daily Fluctuations in Self-Control Demands and Alcohol Intake. Psychology of Addictive Behaviors, 19 (2). 140. DOI: https://doi.org/10.1037/0893-164X.19.2.140
[39] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari and Susan A Murphy. 2016. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Annals of Behavioral Medicine. 1-17.
[40] Inbal Nahum-Shani, Shawna N Smith, Ambuj Tewari, Katie Witkiewitz, Linda M Collins, Bonnie Spring and S Murphy. 2014. Just in time adaptive interventions (jitais): An organizing framework for ongoing health behavior support. Methodology Center technical report (14-126).
[41] Clayton Neighbors, Mary E Larimer and Melissa A Lewis. 2004. Targeting misperceptions of descriptive drinking norms: efficacy of a computer-delivered personalized normative feedback intervention. Journal of consulting and clinical psychology, 72 (3). 434.
[42] Clayton Neighbors, Christine M Lee, Melissa A Lewis, Nicole Fossos and Mary E Larimer. 2007. Are social norms the best predictor of outcomes among heavy-drinking college students? Journal of studies on alcohol and drugs, 68 (4). 556-565.
[43] NIAAA NIH, Drinking Levels Defined | National Institute on Alcohol Abuse and Alcoholism (NIAAA). Accessed from https://www.ncbi.nlm.nih.gov/pubmed/
[44] NIAAA NIH, What's at-risk or heavy drinking? - Rethinking Drinking - NIAAA. Accessed from https://www.ncbi.nlm.nih.gov/pubmed/
[45] Jeremy Northcote and Michael Livingston. 2011. Accuracy of self-reported drinking: observational verification of 'last occasion'drink estimates of young adults. Alcohol and Alcoholism, 46 (6). 709-713.
[46] Josephine Palmeri. 2011. Peer Pressure and Alcohol Use Amongst College Students. Online Publication of Undergraduate Studies, New York University. Accessed on February, 24. 2014.
[47] David Martin Powers. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.
[48] Jürgen Rehm, Thomas K Greenfield, Gordon Walsh, Xiaodi Xie, Linda Robson and Eric Single. 1999. Assessment methods for alcohol consumption, prevalence of high risk drinking and harm: a sensitivity analysis. International Journal of Epidemiology, 28 (2). 219-224.
[49] Nazir Saleheen, Amin Ahsan Ali, Syed Monowar Hossain, Hillol Sarker, Soujanya Chatterjee, Benjamin Marlin, Emre Ertin, Mustafa al'Absi and Santosh Kumar. Year. puffMarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 999-1010.
[50] Patricia Motos Sellés, María Teresa Cortés Tomás, José Antonio Giménez Costa and Fernando Cadaveira Mahía. 2015. Predictors of weekly alcohol drinking and alcohol-related problems in binge-drinking undergraduates. Adicciones, 27 (2).
[51] Saul Shiffman. 2009. Ecological momentary assessment (EMA) in studies of substance use. Psychological Assessment, 21 (4). 486-497. DOI: https://doi.org/10.1037/a0017074
[52] Muhammad Shoaib, Stephan Bosch, Hans Scholten, Paul J. M. Havinga and Ozlem Durmaz Incel. 2015. Towards detection of bad habits by fusing smartphone and smartwatch sensors. In International Conference on Pervasive Computing and Communication Workshops, IEEE, 591-596. DOI: https://doi.org/10.1109/PERCOMW.2015.7134104
[53] Muhammad Shoaib, Stephan Bosch, Hans Scholten, Paul JM Havinga and Ozlem Durmaz Incel. Year. Towards detection of bad habits by fusing smartphone and smartwatch sensors. In Pervasive Computing and Communication Workshops (PerCom Workshops), 2015 IEEE International Conference on, IEEE, 591-596.
[54] Jeffrey S Simons, Thomas A Wills, Noah N Emery and Russell M Marks. 2015. Quantifying alcohol consumption: self-report, transdermal assessment, and prediction of dependence symptoms. Addictive behaviors, 50. 205-212.
[55] Joanne R Smith, Winnifred R Louis, P Wesley Schultz, Clayton Neighbors, Megan Jensen, Judy Tidwell, Theresa Walter, Nicole Fossos and Melissa A Lewis. 2011. Social-norms interventions for light and nondrinking students. Group Processes & Intergroup Relations, 14 (5). 651-669.
[56] Joshua M Smyth and Kristin E Heron. 2016. Is providing mobile interventions "just-in-time" helpful? An experimental proof of concept study of just-in-time intervention for stress management. In Wireless Health.
[57] Joshua M. Smyth and Kristin E. Heron. Year. Is providing mobile interventions "just-in-time" helpful? an experimental proof of concept study of just-in-time intervention for stress management. In IEEE Wireless Health (WH), 1-7. DOI: https://doi.org/10.1109/WH.2016.7764561
[58] Brian Suffoletto, Clifton Callaway, Jeff Kristan, Kevin Kraemer and Duncan B Clark. 2012. Text-message-based drinking assessments and brief interventions for young adults discharged from the emergency department. Alcoholism: Clinical and Experimental Research, 36 (3). 552-560.
[59] Brian Suffoletto, Clifton W Callaway, Jeffrey Kristan, Peter Monti and Duncan B Clark. 2013. Mobile phone text message intervention to reduce binge drinking among young adults: study protocol for a randomized controlled trial. Trials, 14 (1). 93.
[60] Brian Suffoletto, Jeffrey Kristan, Clifton Callaway, Kevin H Kim, Tammy Chung, Peter M Monti and Duncan B Clark. 2014. A text message alcohol intervention for young adult emergency department patients: a randomized clinical trial. Annals of emergency medicine, 64 (6). 664-672. e664.
[61] Brian Suffoletto, Jeffrey Kristan, Tammy Chung, Kwonho Jeong, Anthony Fabio, Peter Monti and Duncan B Clark. 2015. An interactive text message intervention to reduce binge drinking in young adults: a randomized controlled trial with 9-month outcomes. PloS one, 10 (11). e0142877.
[62] Brian Suffoletto, Jeffrey Kristan, Laurel Person Mecca, Tammy Chung and Duncan B Clark. 2016. Optimizing a Text Message Intervention to Reduce Heavy Drinking in Young Adults: Focus Group Findings. JMIR Mhealth Uhealth, 4 (2). e73. DOI: https://doi.org/10.2196/mhealth.5330
[63] Rob Turrisi, James Jaccard, Racheal Taki, Heather Dunnam and Jennifer Grimes. 2001. Examination of the short-term efficacy of a parent intervention to reduce college student drinking tendencies. Psychology of Addictive Behaviors, 15 (4). 366.
[64] Rui Wang, Min SH Aung, Saeed Abdullah, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill and Emily A Scherer. Year. CrossCheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 886-897.
[65] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou and Andrew T Campbell. Year. SmartGPA: how smartphones can assess and predict academic performance of college students. In Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing, ACM, 295-306.
[66] Elissa R Weitzman, Toben F Nelson and Henry Wechsler. 2003. Taking up binge drinking in college: The influences of person, social group, and environment. Journal of Adolescent Health, 32 (1). 26-35
[67] Vik, Peter W., Kayleen A. Culbertson, and Kristie Sellers. 2000. Readiness to change drinking among heavy-drinking college students. Journal of Studies on Alcohol, 61, 674-680.
[68] Maggs, Jennifer L., Lela Rankin Williams, and Christine M. Lee. 2011. Ups and downs of alcohol use among first-year college students: Number of drinks, heavy drinking, and stumble and pass out drinking days. Addictive Behaviors, 36, 197-202.

Received November 2016; revised February 2017; accepted March 2017