Learning to Detect Heavy Drinking Episodes Using Smartphone Accelerometer Data

Jackson A Killian[1] , **Kevin M Passino**[2] , **Arnab Nandi**[2] , **Danielle R Madden**[1] , **John Clapp**[1]
[1]University of Southern California, Los Angeles CA 90089, USA
[2]The Ohio State University, Columbus OH 43210, USA
{jakillia,dmadden,johnclap}@usc.edu,{passino.1,nandi.9}@osu.edu

Abstract

Excessive alcohol consumption is a significant cause of death worldwide and an especially severe risk on college campuses. Recent work aimed at promoting healthier drinking habits has shown promise for the effectiveness of just-in-time adaptive interventions (JITAIs) delivered on mobile platforms *just before* the onset of heavy drinking episodes. However, delivering well-timed JITAIs is difficult for alcohol-related interventions because accurately detecting the onset of such episodes is challenging. Recent work has explored how smartphone data can be used to classify user drinking behavior, but current methods lack generalizability or make liberal use of private user information. We address these shortcomings to develop a reliable mobile classifier that uses only non-sensitive accelerometer data to detect periods of heavy drinking. Additionally, we examine multiple models and discern a new feature set that increases prediction power by as much as 14%. To build our data set, we collected and analyzed smartphone accelerometer readings and transdermal alcohol content (TAC) for 13 subjects participating in an alcohol consumption field study. The TAC readings served as the ground-truth when training the system to make classifications, unlike previous literature which used potentially biased self-reports. Our best classifier detected heavy drinking events with 77.5% accuracy. The deidentified dataset is available on the UCI Machine Learning Repository.[1]

1 Introduction

Excessive alcohol consumption is an avoidable health risk, yet in 2016 it accounted for 5.3% of deaths worldwide [WHO, 2018]. On college campuses, alcohol-related risk is especially dramatic due to higher rates of heavy alcohol use [SAMHSA, 2015]. Thus, social workers have studied how to reduce heavy drinking habits in college students through interventions such as education programs [Brown-Rice *et al.*, 2015], motivational feedback [Borsari and Carey, 2000], and social media campaigns [Thompson *et al.*, 2013] to name a few. With the advent of mobile technologies, researchers have recently begun to investigate the effectiveness of mobile interventions. One study showed that weekly mobile-based interventions can be effective in reducing alcohol consumption in students [Suffoletto *et al.*, 2015], suggesting that students are receptive to mobile communication about drinking. However, a recent study which delivered *hourly* mobile interventions to participants during drinking events showed no significant reduction in the amount of alcohol consumed [Wright *et al.*, 2018], suggesting that overly frequent messaging can reduce the effectiveness of interventions. This highlights the need for accurate, *targeted* messages to participants during drinking episodes. In fact, such just-in-time adaptive interventions (JITAIs) are an active and promising area of research for health domains such as physical inactivity [Consolvo *et al.*, 2008], smoking [Riley *et al.*, 2008], obesity [Patrick *et al.*, 2009], and alcoholism [Nahum-Shani *et al.*, 2017]. One study of recovering alcoholics showed that JITAIs delivered while approaching a bar significantly reduced risky drinking behavior [Gustafson *et al.*, 2014], showcasing how well-timed messages delivered *just before* risky episodes could promote healthier behavior. While promising, work is needed to design JITAIs that apply to college students in general, since their drinking episodes can begin in a variety of complex scenarios from bars, to house parties, to private settings.

The most reliable method for detecting a drinking event is by directly measuring blood alcohol content (BAC) or a proxy such as transdermal alcohol content (TAC). To deliver alcohol-related JITAIs, researchers must passively measure BAC or TAC in real time, but this can be challenging. Some smartphone applications allow users to enter their height, weight, and number of drinks consumed over a period of time to calculate their estimated BAC, but these require active user input that could lead to selection bias and hinder large-scale adoption [Myrecek, 2019]. Some smartwatches can measure TAC but these devices are expensive [BACtrack, 2019] among other roadblocks [Adapa *et al.*, 2018]. *In this work we develop a smartphone-based system to passively track a user's level of intoxication via accelerometer signals to support the delivery of mobile just-in-time adaptive interventions during heavy drinking events.* Smartphone-based solutions are readily scalable since they require no new technological adoption by the user. Further, we ensure that our system minimizes the chance that users will become annoyed or uncomfortable such that they disengage with the system entirely for the following two reasons. First, our system's passive nature requires no user action beyond their normal behavior to generate measurements. Second,

[1]https://archive.ics.uci.edu/ml/datasets/Bar+Crawl%3A+Detecting+Heavy+Drinking

our system uses only raw accelerometer readings rather than highly sensitive user data such as keystrokes, calls, or location. Minimizing the use of sensitive data is of paramount importance for the system's adoption as digital privacy concerns grow. In this work, we make three key contributions as follows.

Sensor-based Field Study: We collected smartphone accelerometer data for 13 students participating in a one-day "bar crawl" event where students, as a group, visited all the bars in a certain region on campus. Further, each student wore an ankle bracelet that measured TAC. To the best of our knowledge, this was the first study to **both** 1) collect data in a field setting and 2) use sensors to measure intoxication rather than self-reports. Thus our findings are 1) applicable to real-world drinking scenarios and 2) our classifications are free from user bias.

Predictive New Features: We adapt Mel frequency cepstral coefficients (MFCC)—traditionally used for sound classification—and apply them to classifying accelerometer data. We show that MFCC covariance features improve results up to **14%**, suggesting these features should be included in future studies classifying accelerometer data.

Heavy Drinking Classifier: We develop a model that makes classifications on 10-second windows of accelerometer data to support the delivery of interventions in real-time. We train several machine learning classifiers including a convolutional neural network, shallow neural network, random forest, and support vector machine to make classifications of sober (TAC < 0.08) vs. intoxicated (TAC ≥ 0.08). The random forest performs the best, achieving an accuracy of **77.5%**.

2 Prior Work

Applying machine learning to classify levels of intoxication using mobile data has recently gained popularity. An early three-participant lab test used smartphone accelerometer readings of users walking on a treadmill to classify intoxication [Kao *et al.*, 2012]. Though they identify useful predictive features, their results are limited by the controlled nature of the lab experiment. Later studies built on this by conducting field tests that could capture the various orientations and manipulations of a mobile phone throughout normal use [Arnold *et al.*, 2015; Gharani *et al.*, 2017]; both trained models on accelerometer data from smartphones in field studies to classify intoxication. However, both used self-reports to measure ground-truth intoxication levels which limits the reliability of their models. Our work stands apart from the previous accelerometer-only classifiers because we collect data from a field test *and* use sensors to establish the ground truth labels ensuring that our model is both generalizeable and reliable.

One related field study used a multitude of mobile data, such as keystroke speed, sent/received calls, location and more to classify sobriety [Bae *et al.*, 2017]. Encouragingly, their model achieved a very high classification accuracy and allowed for theoretical intervention within a half-hour. However, the accuracy of their model came at the cost of constantly sensing and storing sensitive personal information. As data privacy concerns grow globally, use of such sensitive information could deter users from using the platform when implemented for campuses at scale. Thus, learning to make accurate classifications with less sensitive data is of paramount importance. Additionally, their ground truth intoxication was established using potentially-biased self-reports. Our work stands apart from this study because we focus on maximizing the predictive power of *only* non-sensitive data and we utilize sensors for ground truth measurements.

3 Data

We gathered data via a field study in which students participated in an annual "Bar Crawl" event. The study design is as follows.

Participant Eligibility. Participants were required to be an active consumer of alcohol (they have consumed alcohol at least once in the past week). Potential participants must have been planning to attend the Senior Bar Crawl on May 2nd, 2017. They were also required to be a current Ohio State student, 21 years of age or older (legal minimum drinking age), single (not married), and own a smart phone (an Android or Apple device). We restricted the study to single (not married) students since that is by far the most common status of all college undergraduate students.

Recruitment. Undergraduate students were recruited through flyers and announcements in campus newspapers and health-related undergraduate newsletters or courses. Those with an interest in participating were instructed to call our office. Participants were screened during these phone calls to determine eligibility. We recorded only the first name and the last initial of participants and scheduled their baseline appointment the morning of the crawl.

Data Collection. We recruited 20 undergraduate students according to the above protocol. The recruited population was made up of 10 men and 10 women, each in their senior year, aged 21-23 (average age of 22.) 17 identified as white, 1 as Latino/Hispanic, 1 as Asian, and 1 as African American. The participants were offered moderate financial compensation to share mobile accelerometer data and wear a sensor for measuring TAC throughout the event. We developed a simple application for iPhone and Android to sample triaxial accelerometer readings at 40Hz, which were periodically sent to an InfluxDB server [InfluxData, 2019]. The TAC data was sampled every 30 minutes using a SCRAM ankle bracelet sensor [Zettl, 2002]. Each participant was verified to have 0.0 TAC when fitted with the SCRAM bracelet on the morning of the event, ensuring that sober data was collected for each participant. After being fit with the bracelets, students then voluntarily engaged in drinking activities throughout the 18-hour event "as normal," i.e. they were given no behavioral instructions from our team. Statistics on the final collected TAC are shown in **Table 1**. Over 30M accelerometer samples were collected and used in final processing. All data was fully anonymized after collection. This study was approved by The Ohio State University Institutional Review Board (2016B0092).

Data Cleaning. Our custom application for collecting accelerometer data installed successfully on all but 1 participant's mobile device. Of the remaining 19 participants, SCRAM reported that 6 of the ankle bracelets produced data that was indicative of a malfunction (power failure, interference from clothing, etc.) The rest of this work, including **Table 1**, uses only the data from the remaining 13 participants.

Both the TAC ankle sensors and smartphone accelerometers were prone to noisy readings due to splashes of alcohol and low sensor-quality respectively. Thus, to smooth each set of time-series data we used MATLAB's signal processing toolbox to implement low-pass filters for removing noise above a given frequency, *fstop*. Namely, we used a Chebyshev Type II filter, which

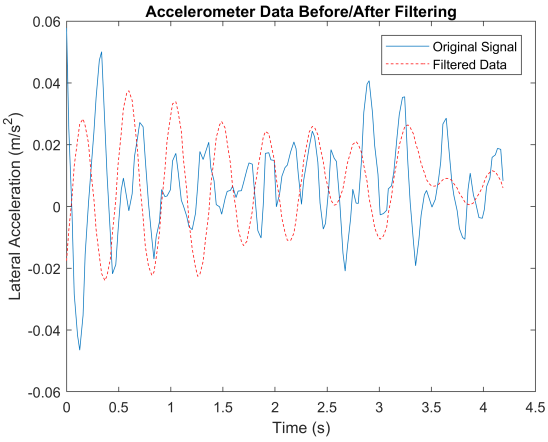
Table 1: Alcohol Consumption Statistics. TAC values are in g/dl where 0.08 is the legal limit for intoxication while driving. Inner quartiles are the 25th, 50th, and 75th quartiles respectively. Time-to-last-drink is calculated as time between TAC sensor initialization and the time of the final local TAC maximum > 0.02. Note that many subjects continued drinking after the conclusion of the event.

Statistic	Value
Mean TAC	0.065 +/- 0.182
Max TAC	0.443
TAC Inner Quartiles	0.002, 0.029, 0.092
Mean Time-to-last-drink	16.1 +/- 6.9 hrs

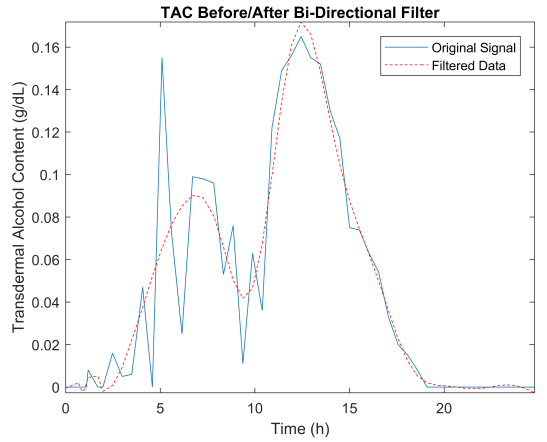
has a steep roll-off at higher orders [MathWorks, 2019]. For the TAC data, we used a 10th order Chebyshev Type II filter and *fstop* = 1e-4Hz. The order and *fstop* were determined empirically to work well. We applied the filter in the forward and reverse direction in order to maintain phase, so that the cleaned TAC could still be matched to corresponding accelerometer readings through their time stamps. This is important since the sampling frequency of the TAC data is over a much larger time scale than that of the motion data, so small phase changes could result in the TAC readings being shifted by several minutes. Then, to account for the time it takes for alcohol to exit the bloodstream and evaporate through the skin, we subtracted 45 minutes from each TAC reading to obtain readings indicative of real-time intoxication [Clapp *et al.*, 2017]. For each axis of accelerometer data we found empirically that a 15th order Chebyshev type II filter worked well and we set *fstop* = 2.7Hz since 1) the average human walking frequency is about 2Hz [Ji and Pachi, 2005] and 2) we found that the large majority of signal was less than 3Hz. Note that for the accelerometer data the filter was only applied in the forward direction since phase changes in that data were on the order of one second and did not affect downstream analysis. Examples of the accelerometer and TAC readings before/after filtering are shown in **Fig. 1**.

Segmentation. Next we segmented each user's stream of 18 hour time-series accelerometer data into windows of a predefined length. Our procedure involved two steps. Over the course of the bar-crawl event, participants' mobile devices experienced occasional periods in which they lost internet connection or battery power. These cases resulted in either zero-readings or a lack of data collection altogether for the duration of the outage. So we first split each participant's stream of data into segments separated by at least two minutes of zero-data or a lack of readings.

Then we tried two approaches. For the first, we intended to isolate segments of accelerometer data collected when a participant was walking. To accomplish this, we used a 4-second-wide sliding window over each segment in which we analyzed the frequency content of the data, keeping windows which were rich in frequency content near 2Hz. Consecutively accepted windows were concatenated up to a maximum length of 60 seconds, resulting in windows of data between 4 and 60 seconds. For our second approach, we simply split each segment into windows of length x, where x ranged from 4 seconds to 2 minutes. Further, each window was required to have at least 90% of the expected data points for a sampling rate of 40Hz.



(a) Accelerometer data



(b) TAC data

Figure 1: A 4-second window of accelerometer data (a) and one student's TAC data (b). Each plot shows the data before (solid) and after (dashed) applying a low-pass filter.

4 Methods

Formulation. The prediction task is as follows: given a sample of accelerometer data, classify the sample as corresponding to TAC *above* or *below* some preset threshold. We chose this binary prediction formulation rather than a regression formulation despite having continuous readings for TAC because the accuracy of readings from the SCRAM sensors in our study varied with TAC level. Specifically, readings from SCRAM sensors are known to be relatively inaccurate for low levels of alcohol consumption, but far better during binge drinking episodes [Barnett *et al.*, 2014]. Preliminary analysis confirmed that classifiers outperformed regression techniques on our data.

Features. In preparing time-series data for classification by machine learning algorithms, it is common to extract features from both the time domain and frequency domain. In both cases, it is also common for each calculated feature to use a two-tiered windowed approach to characterize the data as it changes with time. That is for a given metric of some, say, 4-second segment, we further segment the window down to 1-second segments. We

then calculate the metric for each small segment, then compute the mean, variance, max, and min of the metric over the 4 smaller segments to characterize how it changes over time. We additionally compute the mean of the lower third and upper third of sorted values, creating a total of 6 summary statistics per metric.

Kao et al. [Kao *et al.*, 2012] found that the length of a step and the time between steps changed as participants became intoxicated. Arnold et al. [Arnold *et al.*, 2015] also explored gait cadence, signal skewness, and signal kurtosis in the time domain, as well as average power and ratio of spectral peaks in the frequency domain. We calculate these and many other features using the two-tiered window approach with PyAudioAnalysis [Tyiannak, 2019] and SciPy [Jones *et al.*, 2001]; the full list is described in the top section of **Table 2.** Next, [Lamoth *et al.*, 2010] showed that the root-mean-square (RMS) of lateral accelerations was significantly different in a study of transfemoral amputees between control and amputee populations. We posited that this quantified a lack of control of one's center of mass which would also be true of our intoxicated subjects. Thus for each sample we calculated the RMS of the signal over each axis. Finally, we drew from a technique used in speech recognition tasks which summarizes the energies of a signal in the frequency domain into 13 bins, known as MFCCs. We calculated MFCCs for given small windows of a segment (i.e. 1 second windows) then calculated the covariance of the resulting matrix of coefficients. For example, for a 4-second segment, we split into 1 second windows then calculated the 13 MFCCs over the 4 window resulting in a matrix M of dimension 13×4. We then calculated the covariance matrix as MM^T. In this way, the covariance matrix of the MFCCs captured the change in frequency content of the signal over time. In our experiments, we flattened the covariance matrix and kept only the entries above the diagonal since the resulting 13×13 matrix was symmetric. We calculated these 91 coefficients for each axis of accelerometer data with itself as well as for each axis with each other (i.e. $X\times X$, $X\times Y$, $X\times Z$, etc.). The full summary of all 1215 features can be found in **Table 2**.

Classifiers. We built and trained four different classifiers, each of which was tuned on the training data via 5-fold grid search. First, we built a shallow Multilayer Perceptron Network (MLP) with TensorFlow [Abadi *et al.*, 2015], setting hidden layer size to 256 and learning rate to 0.001. Next we built an SVM with a radial basis function, using LIBSVM [Chang and Lin, 2011]. We set C and γ as 0 and 2 respectively. Then we built a random forest using Python's Scikit-Learn [Pedregosa *et al.*, 2011] using a forest with 700 trees.

Finally, we built a convolutional neural network (CNN) due to their success in other time-series classification tasks such as instrument classification [Park and Lee, 2015; Han *et al.*, 2017] and activity recognition [Zhang *et al.*, 2015; Jiang and Yin, 2015]. In general, these approaches first convert the signal to the frequency domain by computing FFTs over several windows of the signal. This, however, leads to very large inputs which prohibited training in our case. Instead, we computed 16 MFCC coefficients for each second of data on each individual axis resulting in far smaller inputs (i.e. 10x16x3) which were much easier to use for training. The architecture was as follows for a $10\times16\times3$ input: CONV (6×32) POOL (2×2) BATCHNORM () FC (1028) DROPOUT (0.5) FC (2). We built the network with Keras [Chollet, 2015] and trained using cross-entropy loss, Adam Optimizer, and a batch size of 256.

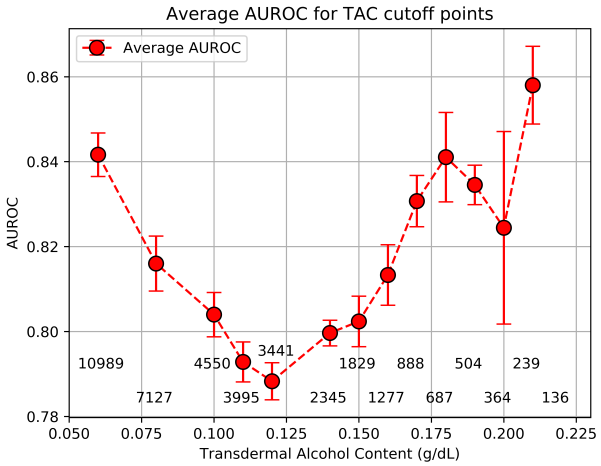


Figure 2: Area under the receiver operating characteristic curve vs. the TAC cutoff point for separating classes. In general, classifier performance is better when the cutoff is moved to extreme points but worse at mid-range points, suggesting that student behavior is somewhat similar for TACs in the range [0.10-0.15]. The number aligned with each marker denotes the number of data points with TAC above that threshold.

Dimensionality Reduction. We generate 1215 features per window of accelerometer data. However, not all of those features were essential for differentiating between classes, and reducing the size of inputs can speed up training. We left the fraction-of-most-important-features-to-keep, λ, as a parameter to tune via grid search. To determine the importance of each feature we trained Scikit-Learn's Random Forest Classifier on randomized subsets of the data, then evaluated the importance of each feature calculated as the normalized reduction of the Gini impurity brought by that feature [Pedregosa *et al.*, 2011].

Parameter Tuning. Three parameters needed tuning regardless of the underlying machine, namely the TAC cutoff between classes, the fraction of important features to keep (λ) and the length of the window of accelerometer data. These were calculated individually using 4-fold grid search on the training data.

We suspected that the differences between "sober" and "intoxicated" classes would become more pronounced and easier to discriminate as we increased the cutoff. However, **Fig. 2** shows that classifier performance followed a U-shaped curve, suggesting that student behaviors are very distinct when either sober or dangerously intoxicated, but tend to have overlap in the middle that can hinder learning. We chose to sacrifice some classification power by keeping the cutoff at 0.08, since we intended our model to be used to deliver JITAIs and since this marks the onset of a binge drinking event as defined by the National Institute on Alcohol Abuse and Alcoholism [NIAAA, 2019].

Fig. 3a shows the performance of our classifier vs. the percent of top most important features kept. Gains in accuracy are made until about 0.2, after which no additional improvements can be seen and training becomes cumbersome. Therefore we set $\lambda = 0.2$. Finally, we posited that our segmentation method to extract only windows of walking data would provide the best downstream results by reducing noise and narrowing the scope of our classifier. However, using this method we obtained a maximal classification

Table 2: Features calculated for each sample of accelerometer data. The number in brackets denotes the number of features generated by that measure. The top section describes features calculated per short-term window. Each short-term feature becomes the basis for 6 summarizing statistics for each of the 3 axes over the full window. Each of the features described in the first 10 rows are calculated per window, then used again to find the difference between the current and previous window resulting in double the features. Features in the bottom section were calculated separately from the two-level window summary technique.

Feature(s)	Definition
Mean [36]	Average of raw signal
Standard Deviation [36]	Standard deviation of raw signal
Median [36]	Median of raw signal
Zero Crossing Rate [36]	Number of times signal changed signs
Max/Min, Raw/Abs [144]	Max/Min of raw/absolute signal (4 total metrics)
Spectral Entropy [72]	Entropy of energy in both the frequency and time domain (2 total metrics)
Spectral Centroid [36]	Weighted mean of frequencies
Spectral Spread [36]	Measure of variance about the centroid
Spectral Flux [36]	Measure of speed of change between two consecutive FFTs
Spectral Roll-off [36]	Frequency under which 90% of energy is contained
Max Frequency [18]	Max frequency from FFT
Gait Stretch [18]	Difference between max and min of one stride
Number of Steps [18]	Total steps taken during a window
Step Time [18]	Average time between two steps
Cadence [18]	Total steps over total time
Skewness [18]	Measure of asymmetry of time-series signal
Kurtosis [18]	Heaviness of tail/Fourth moment of time-series signal
Average Power [18]	Average power over a Welch's power spectrum distribution
Spectral Peak Ratio [18]	Ratio of largest peak to second-largest peak
RMS [3]	Root-mean-square of accelerations for each axis
MFCC Covariance [546]	MFCC covariance entries for each of 6 axis combinations

accuracy of 65% (16% worse than our best result.) Thus we used simply cut windows of non-zero accelerometer data. In general with this approach we suspected that an ever-increasing window length would add more information to be captured by our features giving us better results downstream. However, **Fig. 3b** demonstrates that the opposite was generally true. We obtained the best performance when using the shortest window of 10 seconds. Encouragingly, this is desirable for real world applications, since the shorter the window length, the less resources that are expended and more likely one is to be able to gather the necessary data to perform a classification.

Table 3: Final classification accuracies, precision, and recall per machine. Best results are bolded. Note that recall also serves as the in-class accuracy for the intoxicated data.

Classifier	Acc.	Sober Acc.	Precision	Recall
MLP	0.7328	0.7561	0.5981	0.6885
CNN	0.7427	0.7892	0.6262	0.6564
SVM	0.7467	0.8124	0.6362	0.6224
RF	**0.7748**	**0.8153**	**0.6658**	**0.6979**

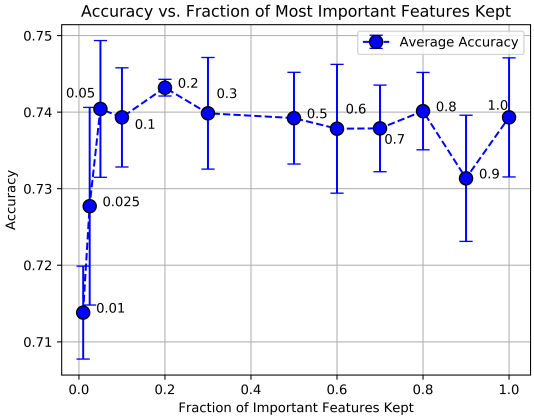
5 Results and Discussion

Using a window length of 10 seconds and $\lambda = 0.2$, we had 26,087 rows of data each with 243 features. Intoxicated samples (TAC > 0.08) made up about one-third of the data. We randomized the data and split 25% as the test set. **Table 3** shows the results for each classifier. The random forest is the clear best performer in all categories. Among the remaining classifiers, notably the shallow MLP network has the worst overall accuracy but is the best at detecting intoxicated samples.

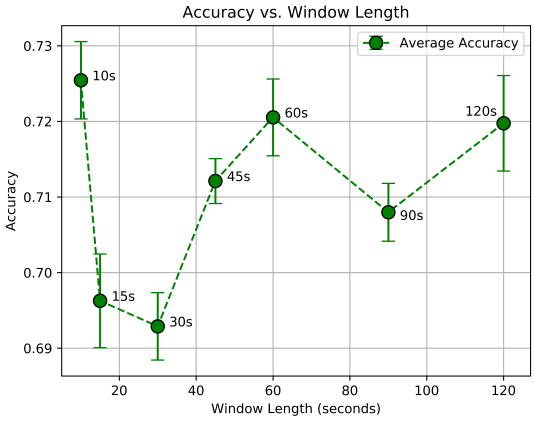
To the best of our knowledge, this is the best-to-date binary classification accuracy on an accelerometer/intoxication task, reaching 77.5%. We were able to achieve this result for two main reasons. First, we accessed a participant cohort that was sufficiently large and that generated data for a time-span long enough to allow us to collect more than 25,000 observations, whereas the classifiers of some comparable studies have had far less training data. Gathering so much data allowed us to train classifiers prepared for many different input cases without suffering from overfitting. Further, we discovered that calculating MFCC covariance matrices as features drastically improved classification accuracy.

Thus, in **Table 4** we quantify the predictive power added by these features by retraining and testing the MLP, RF, and SVM on the data without the MFCC covariance features. Note that this analysis did not apply to the CNN because it did not use covariance entries. Since this significantly changed the number of features, before retraining we re-tuned the parameters for each classifier. Grid search showed that $\lambda = 0.2$ was still the best fraction of important features to keep for all machines. MLP and RF parameters were the same as the previous step, but for the SVM we set $C = 8$ and $\gamma = 4$.

Finally, since the random forest performed the best across all



(a) Fraction of features to use



(b) Window length

Figure 3: Grid search run to determine the best (a) fraction of the most important features to keep and (b) seconds of consecutive accelerometer data to consider (window length.)

Table 4: Additional Accuracy from MFCC Features

Classifier	w/ MFCC	w/o MFCC	Change
MLP	0.7328	0.6581	0.1135
SVM	0.7467	0.6697	0.1150
RF	0.7748	0.6763	**0.1456**

categories, we further investigated its robustness by exploring the variance of its errant calls. **Fig. 4** shows a box plot of the actual TAC of samples that our model misclassified. The plot demonstrates that the variability of even our most robust machine is still high; especially for intoxicated samples. For sober samples that we wrongly called intoxicated (false positives), 50% of samples had TAC between 0.06 and 0.08, which is reasonable. However, 25% of false positive calls occurred for data between 0.01 and 0.04 TAC which is quite distant from the legal limit. For intoxicated samples that we wrongly classified as sober (false negatives), the variance of our machine is even more drastic. Specifically,

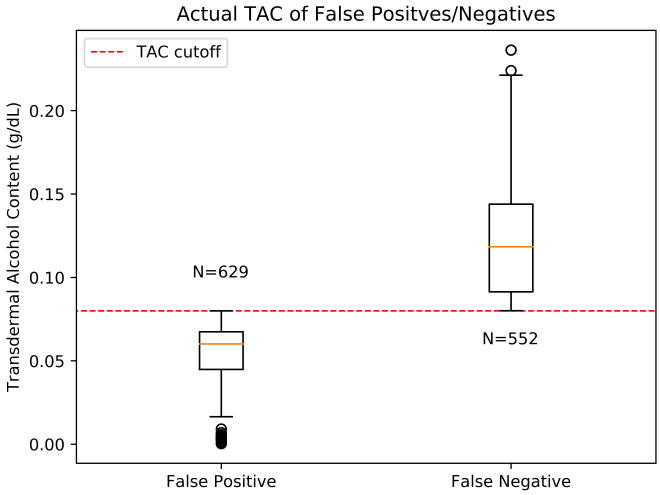


Figure 4: A boxplot showing the actual TAC for samples that were misclassified by the random forest. The left shows the distribution for false positive cases and the right for false negative cases. The red dashed line denotes the TAC cutoff of 0.08.

25% of these samples had TAC of $0.14 - 0.23$ where patients would be considered extremely intoxicated, but our classifier missed them. We take this opportunity to highlight that, though our system takes a step in the right direction for using mobile phones to classify sobriety, more work would be needed to use our tool in situations with social consequences (i.e. law enforcement.)

Despite its variance, our model is well-suited to help develop JITAIs for heavy drinking scenarios. Our model is capable of making classifications of sobriety every 10 seconds. Thus, one could use the model to make several informed predictions over the course of 1 or 2 minutes, then use the whole of the predictions to establish confidence in the user's sobriety. Researchers could then set their own confidence threshold on which to deliver the JITAI within only minutes of the user becoming at risk for a heavy drinking scenario. Further, since our model uses only non-sensitive accelerometer data, *our model could be used for the real-world implementation of JITAIs outside of a research setting since users will not face privacy concerns that might deter them from using the technology.*

6 Future Work and Improvements

Though we achieved a promising level of accuracy, we faced three important challenges. First, while not controlling for phone placement (pocket, purse, etc.) allowed us to capture more general use-cases, it likely hindered our classifier's performance. Including data from gyroscopes might help alleviate this issue and while still maintaining anonymity. However, even with more sensor data more work is needed to incorporate device orientation before classifications will improve. Further, we captured arbitrary user gestures for which it may be difficult to learn general classification rules. This could be remedied by focusing only on one type of gesture such as a walking event. We attempted to extract walking events during post-processing, but found that this generally harmed classification accuracy. Using a mobile device's built-in algorithm for flagging walking data at sensor-time could

be a promising direction. Finally, it is also important to note that all of our classifiers had a higher accuracy for sober data than intoxicated data and that the variance of our best classifier was high for intoxicated subjects. This suggests that more complex techniques or features capable of modeling the diverse set of user actions in heavy drinking scenarios may be needed.

7 Conclusion

We gathered a high-quality dataset for training a machine learning classifier to differentiate between a sober and intoxicated subject using only tri-axial accelerometer signals. The dataset was more reliable than most studies of its nature, given that ground truth intoxication levels were established using sensors, rather than potentially biased self-reports. Further, on our dataset we achieved the highest known accuracy for accelerometer-only binary classification of sobriety, obtaining a test accuracy of 77.5% with a random forest. We also identified a highly informative new set of features to be investigated in future studies; namely MFCC covariance matrix entries which improved classification accuracy by as much as 14%. More work is needed to better understand what differentiates sober from intoxicated accelerometer signals—but these results will serve to improve the baseline for future studies addressing this issue. To support continued research, we have made the de-identified accelerometer and transdermal alcohol content data freely available on the UCI Machine Learning Repository.[2]

Acknowledgments

This work was funded by a seed grant from the College of Social Work at The Ohio State University. We would like to thank The Ohio State University for providing financial support and Brent Leonard and the Ohio AMS Company for their generous assistance with the transdermal alcohol devices utilized in this study.

References

[Abadi *et al.*, 2015] Martín Abadi, Ashish Agarwal, Paul Barham, and et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[Adapa *et al.*, 2018] Apurva Adapa, Fiona Fui-Hoon Nah, Richard H Hall, Keng Siau, and Samuel N Smith. Factors influencing the adoption of smart wearable devices. *International Journal of Human–Computer Interaction*, 34(5):399–409, 2018.

[Arnold *et al.*, 2015] Zachary Arnold, Danielle LaRose, and Emmanuel Agu. Smartphone inference of alcohol consumption levels from gait. In *Healthcare Informatics (ICHI), 2015 International Conference on*, pages 417–426. IEEE, 2015.

[BACtrack, 2019] BACtrack. Bactrack skyn. https://www.bactrack.com/pages/bactrack-skyn-wearable-alcohol-monitor, 2019.

[Bae *et al.*, 2017] Sangwon Bae, Denzil Ferreira, Brian Suffoletto, Juan C Puyana, Ryan Kurtz, Tammy Chung, and Anind K Dey. Detecting drinking episodes in young adults using smartphone-based sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):5, 2017.

[Barnett *et al.*, 2014] Nancy P Barnett, EB Meade, and Tiffany R Glynn. Predictors of detection of alcohol use episodes using a transdermal alcohol sensor. *Experimental and clinical psychopharmacology*, 22(1):86, 2014.

[Borsari and Carey, 2000] Brian Borsari and Kate B Carey. Effects of a brief motivational intervention with college student drinkers. *Journal of consulting and clinical psychology*, 68(4):728, 2000.

[Brown-Rice *et al.*, 2015] Kathleen A Brown-Rice, Susan Furr, and Maribeth Jorgensen. Analyzing greek members alcohol consumption by gender and the impact of alcohol education interventions. *Journal of Alcohol and Drug Education*, 59(1):19–38, 2015.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[Chollet, 2015] François Chollet. Keras. https://keras.io, 2015.

[Clapp *et al.*, 2017] John D Clapp, Danielle R Madden, Douglas D Mooney, and Kristin E Dahlquist. Examining the social ecology of a bar-crawl: An exploratory pilot study. *PLoS one*, 12(9):e0185238, 2017.

[Consolvo *et al.*, 2008] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1797–1806. ACM, 2008.

[Gharani *et al.*, 2017] Pedram Gharani, Brian Suffoletto, Tammy Chung, and Hassan A Karimi. An artificial neural network for movement pattern analysis to estimate blood alcohol content level. *Sensors*, 17(12):2897, 2017.

[Gustafson *et al.*, 2014] David H Gustafson, Fiona M McTavish, Ming-Yuan Chih, Amy K Atwood, Roberta A Johnson, Michael G Boyle, Michael S Levy, Hilary Driscoll, Steven M Chisholm, Lisa Dillenburg, et al. A smartphone application to support recovery from alcoholism: a randomized clinical trial. *JAMA psychiatry*, 71(5):566–572, 2014.

[Han *et al.*, 2017] Yoonchang Han, Jaehun Kim, Kyogu Lee, Yoonchang Han, Jaehun Kim, and Kyogu Lee. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(1):208–221, 2017.

[InfluxData, 2019] InfluxData. Influxdb 1.7 documentation. https://docs.influxdata.com/influxdb/v1.7/, 2019.

[Ji and Pachi, 2005] Tianjian Ji and A Pachi. Frequency and velocity of people walking. *Structural Engineer*, 84(3):36–40, 2005.

[Jiang and Yin, 2015] Wenchao Jiang and Zhaozheng Yin. Human activity recognition using wearable sensors by

[2] https://archive.ics.uci.edu/ml/datasets/Bar+Crawl%3A+Detecting+Heavy+Drinking

deep convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1307–1310. ACM, 2015.

[Jones *et al.*, 2001] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python. http://www.scipy.org/, 2001.

[Kao *et al.*, 2012] Hsin-Liu Cindy Kao, Bo-Jhang Ho, Allan C Lin, and Hao-Hua Chu. Phone-based gait analysis to detect alcohol usage. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 661–662. ACM, 2012.

[Lamoth *et al.*, 2010] Claudine JC Lamoth, Erik Ainsworth, Wojtek Polomski, and Han Houdijk. Variability and stability analysis of walking of transfemoral amputees. *Medical engineering & physics*, 32(9):1009–1014, 2010.

[MathWorks, 2019] MathWorks. Documentation: cheby2. https://www.mathworks.com/help/signal/ref/cheby2.html, 2019.

[Myrecek, 2019] Myrecek. Alcodroid alcohol tracker. https://play.google.com/store/apps/details?id=org.M.alcodroid&hl=en_US, 2019.

[Nahum-Shani *et al.*, 2017] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462, 2017.

[NIAAA, 2019] NIAAA. Drinking levels defined. https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/moderate-binge-drinking, 2019.

[Park and Lee, 2015] Taejin Park and Taejin Lee. Musical instrument sound classification with deep convolutional neural network using feature fusion approach. *arXiv preprint arXiv:1512.07370*, 2015.

[Patrick *et al.*, 2009] Kevin Patrick, Fred Raab, Marc A Adams, Lindsay Dillon, Marian Zabinski, Cheryl L Rock, William G Griswold, and Gregory J Norman. A text message–based intervention for weight loss: randomized controlled trial. *Journal of medical Internet research*, 11(1), 2009.

[Pedregosa *et al.*, 2011] F. Pedregosa, G. Varoquaux, A. Gramfort, and et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Riley *et al.*, 2008] William Riley, Jami Obermayer, and Jersino Jean-Mary. Internet and mobile phone text messaging intervention for college smokers. *Journal of American College Health*, 57(2):245–248, 2008.

[SAMHSA, 2015] SAMHSA. 2015 national survey on drug use and health (nsduh). table 6.84b-tobacco product and alcohol use in past month among persons aged 18 to 22, by college enrollment status: Percentages, 2014 and 2015. https://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabs-2015/NSDUH-DetTabs-2015/NSDUH-DetTabs-2015.htm#tab6-84b, 2015.

[Suffoletto *et al.*, 2015] Brian Suffoletto, Jeffrey Kristan, Tammy Chung, Kwonho Jeong, Anthony Fabio, Peter Monti, and Duncan B Clark. 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, 2015.

[Thompson *et al.*, 2013] Erika Beseler Thompson, Frank Heley, Laura Oster-Aaland, Sherri Nordstrom Stastny, and Elizabeth Crisp Crawford. The impact of a student-driven social marketing campaign on college student alcohol-related beliefs and behaviors. *Social Marketing Quarterly*, 19(1):52–64, 2013.

[Tyiannak, 2019] Tyiannak. Python audio analysis library: Feature extraction, classification, segmentation and applications. https://github.com/tyiannak/pyAudioAnalysis, 2019.

[WHO, 2018] WHO. *Global status report on alcohol and health, 2018*. World Health Organization, 2018.

[Wright *et al.*, 2018] Cassandra Wright, Paul M Dietze, Paul A Agius, Emmanuel Kuntsche, Michael Livingston, Oliver C Black, Robin Room, Margaret Hellard, and Megan SC Lim. Mobile phone-based ecological momentary intervention to reduce young adults' alcohol use in the event: A three-armed randomized controlled trial. *JMIR mHealth and uHealth*, 6(7):e149, 2018.

[Zettl, 2002] J. Robert Zettl. The determination of blood alcohol concentration by transdermal measurement. https://www.scramsystems.com/images/uploads/general/research/the-determination-of-blood-alcohol-concentration-by-transdermal-measurement.pdf, 2002.

[Zhang *et al.*, 2015] Licheng Zhang, Xihong Wu, and Dingsheng Luo. Recognizing human activities from raw accelerometer data using deep neural networks. In *Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on*, pages 865–870. IEEE, 2015.