# Identifying Smoking from Smartphone Sensor Data and Multivariate Hidden Markov Models

Yang Qin, Weicheng Qian, Narjes Shojaati, and Nathaniel Osgood

University of Saskatchewan,
first.last@usask.ca

**Abstract.** Smoking is one of the foremost public health threats listed by the World Health Organization, and surveillance is a key to informing effective policies. High smartphone penetration and mature smartphone sensor data collecting techniques make smartphone sensor data based smoking monitoring viable, yet an effective classification algorithm remains elusive. In this paper, we sought to classify smoking using multivariate Hidden Markov models (HMMs) informed by binned time-series of transformed sensor data collected with smartphone-based Wi-Fi, GPS, and accelerometer sensors. Our model is trained on smartphone sensor time series data labeled with self-reported smoking periods. Two-fold cross-validation shows $A_z$ (area under receiver operating characteristic curve) for HMMs using five features = (0.52, 0.84). Comparison of univariate HMMs and multivariate HMMs, suggests a high accuracy of multivariate HMMs for smoking periods classification.

**Keywords:** Hidden Markov model, smartphone sensor data, tobacco, smoking monitoring

## 1 Introduction

Smoking is one of the biggest public health threats listed by the World Health Organization. Effective tools for smoking recognition can ensure public health surveillance for policy making [11], provide early detection before addiction [3], and aid former smokers to avoid relapse. Detection of smoking status has long relied on biomedical assays based around detection of substances such as cotinine, nicotine [8], carbon monoxide [6] and respiration [2]. Application of such assays normally requires mildly to moderately invasive measurements, from breath tests to provision of saliva to clippings of hair, and many test results are available only after delays measured in days or more.

Studies using wearable sensors for recognition of smoking [9, 1] have showed potentials to avoid the invasiveness of measurements and delays of test results to perform seamless online detection. These studies predominately based on hand-to-mouth gestures and breathing pattern and using specialized hardware to collect data, which can be costly and hard to comply continuously, informativeness from other aspects correlates of smoking have not yet been considered, such as presence outdoors (as required by regional regulations) or designated smoking

areas, activity levels, and characteristic length of dwelling period correlated to the burning time of cigarettes.

In recent years, and paralleling their rapid penetration across diverse strata of society worldwide, smartphone have become an attractive platform for sensor-based data collection on human behaviour. The use of such techniques has been enhanced by the growing maturity of data collection apps (such as the iEpi system [5], UPenn's DREAM project) that make smartphone sensor data a highly available and easily accessible data source for many studies [7]. Feasibility studies on using accelerometer sensor to detect smoking behaviors has been initiated [10], but published studies on fusing sensor data available on smartphone remains absent.

In this paper, we fused various types of sensor data commonly available on smartphone, after considering data completeness, accuracy and informativeness, examined the effects of five transformations for GPS, Wi-Fi and accelerometer sensor data, and applied multivariate hidden Markov model (HMMs) to classify periods to recognize whether smoking was taking place. Finally, we investigated the performance of univariate HMMs and multivariate HMMs, and the impact of tailoring the training and test set to preserve entire smoking cycles.

## 2 Data Processing and Algorithm

**Dataset Description** Data used in the project came from a previously conducted Behavioural Ethics Board-approved study that collected multiple types of sensor data together with self-reported ground truth on smoking behavior by four participants (one did not complete) who were self-reported smokers. The dataset contains labeled data on segments of intervals of smoking and non-smoking periods. The sensor data was collected with a five-minute duty cycle by Ethica system [4, 5] for three participants over one month from April 04, 2015 to May 12, 2015. There are 36 million records from gyroscope, 0.3 million records for location, 1.9 million records for Wi-Fi and 36 million records for accelerometer.

**Data Processing** We found each participant has an extremely long smoking period at the end of their self-labeled smoking-nonsmoking periods, which are apparently outliers and therefore we preserved only a period at the head of each of those extremely long periods, whose length equals to the average length of previous smoking periods of this very person, and trimmed the rest smoking periods at the end.

For accelerometer data, we applied a high-pass filter, using a standard deviation of norms of readings on X, Y, Z accelerometer axes in the 30s timeslot, to separate out the dominant invariant gravitational component. For Wi-Fi data, Received signal strength indication (RSSI) in 30-second timeslots was considered. ECDF of the counts of unique MAC address during 30s timeslots showed better difference between two states than that of maximum RSSI. Maximum RSSI indicated the strongest Wi-Fi signal, and counts of unique MAC address

represented number of accessible networks for the smartphone. The source of location data, either GPS (using satellite) or network (using cell tower and Wi-Fi based location), indicates whether the participant is indoor or outdoor. So the count of GPS readings across 30s timeslots specifically drawn from satellite (as opposed to network) sources was used.

**Multivariate HMM** Using the transformed data described above, a multivariate HMMs was employed to classify smoking and non-smoking intervals based on real world labeled observations. In this model, each state has multiple observations corresponding to readings from Wi-Fi, accelerometer and GPS sensors. The probabilities of observations follow empirical distributions, and observations are assumed to be independent from each other, conditional on being in a given state. So the likelihood of observing a given vector of observed quantities was approximated as the product of independent probability density functions as given by kernel density estimates.

**Two-fold Cross Validation** Hidden Markov models expect sequential observations, therefore when choosing training set and test set, we can not simply sample at random time intervals from data sequence, but rather need to divide the data sequence into disjoint contiguous sequences. Firstly, we made use of a two-fold cross-validation approach, where we cut the sequence from the head to 50%, 55%, 60%, 65%, 70% and 75% of the sequence to ensure sequential observations (including NAs) for training, and used the balance of the observations for testing. Second, We swapped the training set and test set in first step to feed the HMMs.

## 3 Results

Structured learning was used in this project. This work was conducted in several phases. In phase 1, maximum RSSI, counts of unique MAC address during 30s timeslots for Wi-Fi, average norms, standard deviation of norms for accelerometer and counts of GPS reading from GPS source in 30s timeslots were considered as a single feature, respectively. Each of the five features was used to train univariate HMMs, which were then evaluated. In phase 2, Multivariate HMMs using three features and five features were trained and evaluated.

The HMMs were found to yield favorable results in the multivariate cases and in univariate cases considering accel sensor data as feature. As shown below, multivariate HMMs exhibit accuracy over 0.9, and an area under ROC curve (AUC) above 0.8 when collected data is representative. The results of HMMs with a single feature are less favorable than those for multivariate HMMs.

### 3.1 Results with single feature

For using average of norm of the accelerometer as a feature, the AUC for training set and test set with different size of training set range from 0.69 to 0.94 and

from 0.60 to 0.79, respectively. And the error rates of the training set and test set range from 0.096 to 0.27 and from 0.057 to 0.357, respectively. For using of standard deviation of norm of the accelerometer as a feature, the AUC for training set and test set with different size of training set range from 0.76 to 0.92 and from 0.63 to 0.86, respectively. And the error rates for the training set and test set range from 0.05 to 0.12 and from 0.06 to 0.2, respectively, as shown in Figure 1.

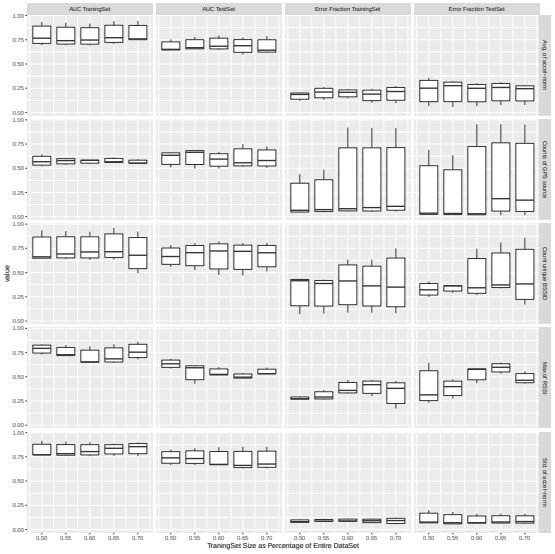


Fig. 1: AUC and error rate of HMMs using avg. and std. of accel-norm, count of GPS source, count of unique BSSID and max of RSSI as single feature

For using readings from Wi-Fi sensor as single feature (maximum RSSI or counts of unique MAC address), for maximum RSSI, the range of AUC for test set is from 0.43 to 0.69, and the range of error rate for test set is from 0.23 to 0.65. For counts of unique MAC address, the range of AUC and error rate for test set ranges from 0.47 to 0.82 and from 0.17 to 0.86, respectively.

In this model, using only one feature derived from the GPS sensor, the range of AUC for training set is from 0.52 to 0.64, for test set is from 0.50 to 0.75. Error rates for training set and test rate are from 0.04 to 0.92, and from 0.015 to 0.96, respectively.

### 3.2 Results with three features

Counts of unique MAC address for Wi-Fi, standard deviation of accel norm and counts of GPS sourced signal were employed together to train the HMMs. The results offer AUC and error rates of the HMMs were shown in Figure 2 . The

range of AUC for the test set is from 0.5 to 0.83 with an average of 0.65, and error rate for test set ranged from 0.017 to 0.96 with average of 0.23.

### 3.3 Results with five features

All five features derived from sensors were employed together to train HMMs. The results for AUC and error rates of the five feature HMMs were also shown in Figure 2. The range of AUC for training set is from 0.5 to 0.98 with average 0.79, and for test set, is from 0.52 to 0.84 with average 0.66. The Wi-Fi and GPS data are location-based features, while the components of the accelerometer data are associated with the body gestures and orientation features of participant. The combination of five features can capture a larger set of information on the current smoking activity of participants. This enlarged information can in turn enhance performance of the HMMs.

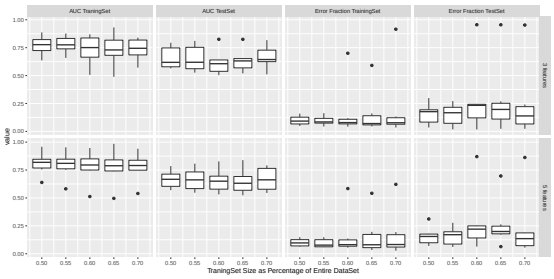


Fig. 2: AUC and error rate of HMMs with three features and five features

## 4 Related Work

Sazonov et al. (2013) developed a wearable sensor system based on hand-to-mouth smoking gestures and breathing pattern [9], Lopez-Meyer et al (2013) further applied a support vector machine and achieved 87% and 80% of average user-independent precision and recall and 90% in user-independent precision and recall [1]. Scholl and Laerhoven (2012) used wearable accelerometer device to collect data and applied basic Gaussian classifier to detect smoking gestures with a precision of 51.2% and 70% of user specific recall [10]. We have not yet found papers about smoking detection using multivariate HMM based on various types of sensor data available from commodity smartphones.

## 5 Limitations and Future Work

Despite high granularity data for each participant, our study is limited by the number of participants, as a future work, we will experiment with a larger participant size. We will also consider covariance among sensor observations for

performance boost and whether commonalities among personal empirical distributions can be extracted and reused on other persons.

## 6 Conclusions

The results of multivariate HMMs demonstrated classification and detection of smoking activity with high accuracy. Compared to single feature HMMs, the multivariate HMMs had higher accuracy, because additional types of sensor data can help us better describe smoking gesture and activity.

This work further suggests that tailoring training set and test set close to entire smoking cycles can improve the performance of HMMs. The large variation in results across participants further raises the possibility that significant components of remaining error rates may be due to limitations in the accuracy of self-reporting of ground truth data on smoking behaviour.

## References

1. Monitoring of cigarette smoking using wearable sensors and support vector machines. IEEE Transactions on Biomedical Engineering 60(7), 1867–1872 (2013)
2. Ali, A., Hossain, S., Hovsepian, K., Rahman, M., Plarre, K., Kumar, S.: mPuff: Automated detection of cigarette smoking puffs from respiration measurements. IPSN'12 - Proceedings of the 11th International Conference on Information Processing in Sensor Networks pp. 269–280 (2012)
3. Community Preventive Services Task Force: Reducing tobacco use and secondhand smoke exposure: mobile phone-based cessation interventions (2013)
4. Ethica Data: https://www.ethicadata.com/
5. Hashemian, M., Knowles, D., Calver, J., Qian, W., Bullock, M.C., Bell, S., Mandryk, R.L., Osgood, N., Stanley, K.G.: iEpi: an end to end solution for collecting, conditioning and utilizing epidemiologically relevant data. In: Proceedings of the 2nd ACM international workshop on Pervasive Wireless Healthcare. pp. 3–8. ACM (2012)
6. Meredith, S.E., Robinson, A., Erb, P., Spieler, C.A., Klugman, N., Dutta, P., Dallery, J.: A mobile-phone-based breath carbon monoxide meter to detect cigarette smoking. Nicotine and Tobacco Research 16(6), 766–773 (2014)
7. Qian, W., Stanley, K.G., Osgood, N.D.: The Impact of Spatial Resolution and Representation on Human Mobility Predictability, pp. 25–40. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
8. Raja, M.: Diagnostic Methods for Detection of Cotinine Level in Tobacco Users: A Review. Journal of Clinical and Diagnostic Research 10(3), 4–6 (2016)
9. Sazonov, E., Lopez-Meyer, P., Tiffany, S.: A wearable sensor system for monitoring cigarette smoking. Journal of studies on alcohol and drugs 74(6), 956–964 (2013)
10. Scholl, P.M., van Laerhoven, K.: A Feasibility Study of Wrist-Worn Accelerometer Based Detection of Smoking Habits. In: 2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. pp. 886–891 (2012)
11. WHO: Tobacco Factsheet (2016), http://www.who.int/mediacentre/factsheets/fs339/