Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 9702730, 11 pages
https://doi.org/10.1155/2018/9702730

Research Article

Investigations on Driver Unique Identification from Smartphone's GPS Data Alone

Arijit Chowdhury,[1] Tapas Chakravarty,[1] Avik Ghose,[1] Tanushree Banerjee,[1] and P. Balamuralidhar[2]

[1] *Tata Consultancy Services, Building 1B, Ecospace, Plot IIF/12, New Town, Rajarhat, Kolkata, West Bengal 700156, India*
[2] *Tata Consultancy Services, Whitefield, Bangalore, India*

Correspondence should be addressed to Tapas Chakravarty; tapas.ieee@gmail.com

Received 28 August 2017; Revised 27 December 2017; Accepted 15 January 2018; Published 8 February 2018

Academic Editor: David Martín

Driver identification is an emerging area of interest in vehicle telematics, automobile control, and insurance. Recent body of works indicates that it may be possible to uniquely identify a driver using multiple dedicated sensors. In this paper, we present an approach for driver identification using smartphone GPS data alone. For our experiments, we collected data from 38 drivers for two months. We quantified the driver's natural style by extracting a set of 137 statistical features from data generated for each completed trip. The analysis shows that, for the "driver identification" problem, an average accuracy of 82.3% is achieved for driver groups of 4-5 drivers. This is comparable to the state of the arts where mostly a multisensor approach has been taken. Further, it is shown that certain behavioral attributes like high driving skill impact identification accuracy. We observe that Random Forest classifier offers the best results. These results have great implications for various stakeholders since the proposed method can identify a driver based on his/her naturalistic driving style which is quantified in terms of statistical parameters extracted from only GPS data.

1. Introduction

Studies on human behavior are gaining importance and a large number of theories are originating on this topic [1]. The area of computational social science is emerging as an important field of study. "Driver profiling" is one such aspect contributing to human behavior models. In recent times, a number of studies have been carried out to extract meaningful and actionable insights from driving data. The primary objective of these studies are to derive an estimate of risk involved in the very nature of driving a land vehicle. Wahlström et al. [2] provide us with a very detailed review of the work done (smartphone based telematics) in the last ten-year span. A large body of the works referred by Wahlström et al. relates to the development of industrial applications like UBI (Usage Based Insurance) and connected car scenarios. However, a larger question remains to be answered: Will the estimation of driving risk be sufficient to model the driving behavior? In an earlier work, Fuller [3] proposed that the drivers try to estimate and maintain a level of task difficulty. Fuller [4] had proposed a TCI (Tasks Capability Interface) model in this regard. This model proposes that the driver is continuously reacting to the demand of the driving task and his/her assessment of own capability. Considering that Fuller's "driver model" is comprehensive, we are then led to believe that each individual will demonstrate uniqueness in driving pattern leading to the possibility of fingerprinting.

Keeping a focus on driver's operative characterization, Lin et al. [5] evaluate different methods such as evaluation of driver's real-time behavior, vehicle state, or monitoring facial expressions. Lin et al. also mentioned that such behavior needs to be classified, before identification. However, the issues related to driver fingerprinting were addressed by Enev et al. [6]. In their paper aptly titled "Automobile Driver Fingerprinting," Enev et al. investigate "the potential to identify individuals" by analyzing their natural driving styles. Here [6], authors demonstrate that, even with a restricted set of sensors, the drivers can be uniquely identified with an accuracy of 87% (99% with top 5 sensors) using just 15 minutes of on-road data. This is a significant advancement in comparison to the work done by Zhang et al. [7] who utilized simulated data [multisession, 20 male drivers] and demonstrated 85% accuracy by using Hidden Markov Model (HMM). Enev et al. also hypothesized that given an availability of enough longitudinal data, "everyone can be distinguished." In another related work on driver identification, Hallac et al. [8] demonstrated that there are unique patterns in individual driving styles which can be detected even for a short drive. Hallac et al. experimentally demonstrated that the vehicle turn signature is often well suited for detecting individual style. In this case, authors designed an experiment consisting of driving a car (several drivers) along a road segment of 150-foot radius. They monitored 12 sensor readings, other than GPS (Global Positioning System). The result shows an average prediction accuracy of 76.9% for two-driver classification and 50.1% for five-driver classification. Authors also mentioned that a fusion approach of prediction models can lead to an enhanced classifier. In a 2015 survey article, Engelbrecht et al. [9] have presented valuable insight in driving behavior analysis. Engelbrecht et al. draw our attention to the fact that smartphone based sensing links the behavior to an individual rather than a vehicle. Further, Engelbrecht et al. note that there are differences between systems which detect driving maneuvers only and those which classify driver behavior. In order to predict future state of maneuvers, it is first necessary to deduce a person's naturalistic driving style. Only then an anomalous behavior can be accurately recognized. In an earlier work, Themann et al. [10] noted that there is a strong need to identify driver's unique preferences and incorporate such models in adaptive cruise control systems (anticipating driving style) so that a largely automated driving can fulfill its promise of improving fuel efficiency.

It is to be noted that there are different objectives for modeling specific driving maneuvers vis-à-vis modeling general driving behavior. The risky sudden maneuvers like hard brake happen at an operational level, in a timeframe of millisec, whereas lane change, turn, and stop are more tactical, on a timescale of seconds. The long term goals of the driver have longer time frames and are strategic in nature. Abuali and Abou-Zeid [11] illustrate the above stated viewpoint with reference to the proposition by Laapotti et al. [12] that there exists a behavioral level on top of the above three levels of hierarchical control model; this layer describes the life skill and general goal of the individual. Thus taking a cue from the above statement, we may predict that identifying a driver's propensities can lead us to understand the person's exercise of choice in different scenarios, whenever she/he is faced with self-assessment of task difficulty and own capability.

Smartphone continues to be a favorite platform for sensing a driver's real-time maneuvers and deriving driving profile. Vlahogianni and Barmpounakis [13] investigate smartphone based analytics for driving behavior assessment. As mentioned by Vlahogianni and Barmpounakis, there is little knowledge about reliability of such smartphone based sensing except for the fact that GPS (in a phone) and accelerometer continue to remain a popular choice. In a recent work Tanprasert et al. have proposed driver identification in real-time using accelerometer and GPS data using unsupervised anomaly detection and neural networks [14]. Acceleration variation is also investigated for the purpose of driver identification in [15] by Phumphuang et al. They also used PCA (Principle Component Analysis) for dimensionality reduction. In another recent work, Junior et al. [16] investigate driver behavior profiling using different Android phone based sensors and different types of machine learning algorithms. In this case, 4 car types and 13 minutes of average drive time are investigated with primary sensors being accelerometer, magnetometer, and gyroscope. In conclusion, authors in [16] mention that Random Forest algorithm "is the best performing MLA with 28 out of 35 best assemblies."

Present authors have worked extensively to improve the reliability of smartphone based GPS measurements [17, 18]. In a recent work [19], we have presented a driver behavior analysis platform where nearly 50000 Kms worth of driving data, obtained from 38 drivers (in their natural environment), is analyzed and statistically modeled with 2D factor analysis, with the factors being aggression and skill. In this work, we utilize the same dataset and investigate much granularity to detect variability in natural driving styles of the individual. We have considered those GPS data as valid where the horizontal accuracy measure is reported to be less than 16 m. A few incomplete datasets are removed for analysis purpose. While majority of collected GPS data (speed, heading, etc.) obtained are found to be quite accurate, we have also applied filters on raw GPS speed measurements (as outlined in [17, 18]) in order to eliminate spurious values. It is often seen that sudden brakes and acceleration lead to momentary loss of data which can lead to wrong estimation of acceleration/deceleration at that instance. However, such instances are few and these have been corrected. It is also felt that a few anomalous events will not affect our proposed aggregate statistical model. In this aggregate model, driver behavior is quantified based on the statistics associated with the completed trip rather than individual events. Also, we have not performed behavior identification with considerations to different road segments separately. Even though we have considered lateral acceleration (derived from GPS data) as an important indicator of driver behavior, our proposed model does not delve deep into specific horizontal curves. Authors in [20] have shown that, along horizontal curves, the perception of the road geometry by drivers affects road safety. Vaiana et al. [20] conducted experiments with 35 participants driving around overall 86 curves with different radii and GPS data was analyzed to understand the driving behavior around the curves. Thus, it is felt that future works may specifically investigate this aspect.

The key contributions of this paper are as follows.

We investigate whether unique identification of driver is possible by using only GPS data measurements. This entails an investigation into the level of accuracy achievable under such circumstances. Our objective is to identify the driver's natural style and distinguish drivers based on their naturalistic driving styles. Towards that goal, feature computation and classification have been performed. Drivers are grouped in groups of 4-5 drivers, based on the proximity

Table 1: Basic statistics and travelled distance for a representative sample of drivers.

Driver ID	Evening trip	Day trip	Morning trip	Distance travelled (Km)	Average speed (m/s)	Total trip duration (minutes)	Total number of trips
D001	121	26	49	3632	11.85	4638	241
D002	20	5	10	365	8.76	902	37
D003	26	1	28	719	8.92	1738	58
D004	30	7	5	884	12.50	1227	48
D005	41	13	26	1057	10.30	2082	103

of regular driving locations as well as trip timings. Such a method is assumed to normalize the environment in which the drivers operate. Then driver identification is studied using different algorithms and accuracy is validated using *k-fold cross-validation*. The results are reported on test sets for different groups. The presented approach enables us to map a journey with the most likely driver taking only GPS data into consideration.

The rest of the paper is organized as follows: in Section 2, we describe our experiment details and the data set generated. Section 3 describes our algorithms and methods applied. In Section 4 we illustrate our results and provide analysis of the same. Finally, we summarize our conclusions in Section 5 of the paper.

2. Experimental Setup and Data Collection

2.1. Experimental Setup. For this study, we collected data from a total of 38 drivers. Majority of the drivers recorded at least 2 trips on weekdays within an observation window of two months. It is to be noted that the authors of this paper played no role in selecting drivers as well as the driving location. However, the subjects form a peer group based on geography and it is assumed that they went about their normal lifestyles, usually travelling from home to work and back, with one or two additional trips. Overall, we have around 4000 individual trips. Our method relies solely on data gathered from smartphone's GPS modules without any knowledge of local road and traffic conditions. Overall a total of 3927 trips are considered from the collected data. Data were collected on weekdays with one or more trips per day. In total, 1233 hours of driving data is captured and the overall car journey length is found to be 50740 Km, covering various regions in the USA. Although the driver's geographical location is available from latitude and longitude, we do not have access to any personally identifiable data such as name, age, gender, car type, and home address for privacy protection. Many users enrolled in this program and were given sequential IDs (D001–D085); we only conducted our study based on 38 drivers who had considerable number of complete trips. As can be observed, the trips are found to be well distributed across different times of the day. A representative sample is presented in Table 1.

Taking long distance trips and short distance trips separated across time in a day creates diversity in runtime condition of the vehicle and thus enriches the data set. Driver D055 was the least frequent traveller with overall 12 trips only. That data is utilized to verify the effectiveness of our algorithm when applied on a relatively low data volume. Although a high identification accuracy for D055 is not expected, we explored the possibility that inclusion of a lower volume data set does not affect the overall result.

2.2. Data Collection and Preprocessing. As mentioned, our primary data collection module is a GPS receiver. The participant driver downloads an app for GPS logging on her smartphone. Each participant is instructed to drive naturally according to his/her preference. Data is collected only when that participant is driving his/her own car (a single vehicle per participant). Smartphone is kept inside car and held fixed with respect to car's body. Since only GPS reading is used, the orientation of smartphone is not important for our work.

In order to evaluate data quality, the knowledge of measurement Precision is of utmost importance [25]. Such Precision is dependent on the quality of GPS receiver of the given smartphone. The integrity and availability of smartphone's GPS signal get affected by events like urban tunneling, sudden change in vehicle speed and direction, and so forth. Handel et al. [25] discuss GPS data integrity enhancement and monitoring in detail. Authors demonstrate that, in order to ensure robust calculation of different figures of merit, "data cleansing and integrity monitoring are much needed." Handel suggests two techniques, namely, second-by-second data and whole trip data. It is stated that a direct differentiation of the GPS speed data will amplify high frequency noises and outliers. A better option will be to clean the speed measurements, by fitting a polynomial model.

For our purpose, we have followed the principle outlined by Handel et al. [25]. To begin with, we eliminate all measurements which offer horizontal accuracy metric of greater than 16 m. Further, we attempt to estimate the true speed (at a given instance) by building a relationship with two immediate past measurements. The method is similar to moving average filter except that the averaging is done with different weight coefficients. The method is illustrated in detail in [17] and is not repeated here. Moreover, it is observed that there are only few missing data points. Since we use aggregate model where each completed trip is taken as a set of data values, an outlier detection algorithm eliminates those which display marked deviation in the statistical properties associated with both lateral and longitudinal acceleration. This aspect is explained in detail in [19]. It is to be noted that the proposed model is quite unlike an event based method. In the present case, a handful of anomalous events are treated as outliers so that we can deduce the natural driving style of the driver.

Thus, the data recorded consists of attributes like timestamp, altitude, course, horizontal accuracy, latitude, longitude, and speed in m/s. All data are collected at 1 Hz rate. The data consists of the following parameters.

(i) Driver ID: a unique ID that distinguishes a driver.

(ii) GPS data: all attributes like speed, location, course, and horizontal accuracy.

Let us assume that v_1, v_2 to v_n be the consecutive speed measurement samples, at time t_1, t_2 to t_n. Similarly, we get θ_1, θ_2 to θ_n as consecutive course (heading) measurements. Let us also assume that a_1, a_2 to a_n be the consecutive discrete acceleration samples (derived from speed measurements) at time t_1, t_2 to t_n, where $\Delta t = t_n - t_{n-1}$ for uniform sampling rate. Then, from these primary GPS measurements, few secondary data are computed. These are as follows:

(i) Longitudinal acceleration:

$$\mathrm{acc}_{\mathrm{Long}} = \frac{(v_n - v_{n-1})}{(t_n - t_{n-1})}. \tag{1}$$

(ii) Angular speed: rate of change of course:

$$\omega = \frac{\partial \theta}{\partial t} * \frac{\pi}{180}. \tag{2}$$

(iii) Lateral acceleration [26]:

$$\mathrm{acc}_{\mathrm{Lat}} = \frac{v^2}{R} = v\omega. \tag{3}$$

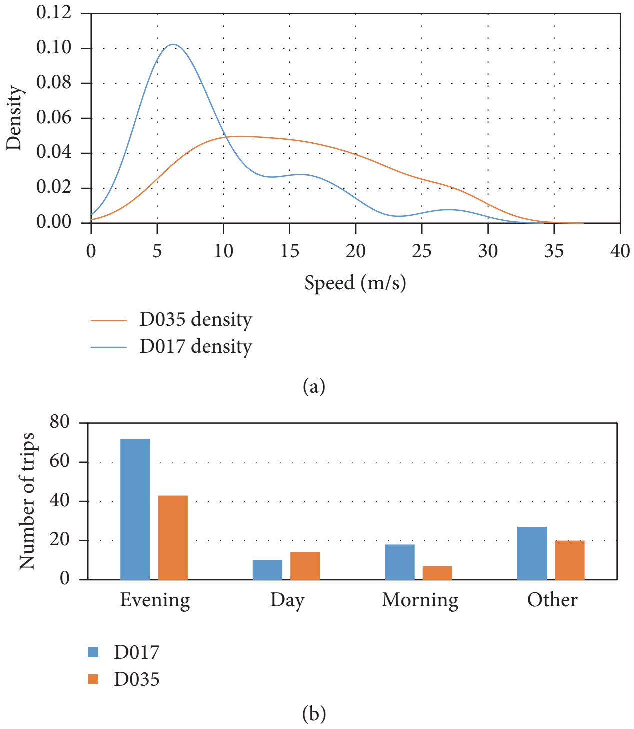
(iv) Jerk: rate of change of acceleration (m/s^3):

$$J_i = \frac{(a_{i+1} - a_i)}{\Delta t} \quad \forall 1 \leq i \leq 3. \tag{4}$$

(v) Jerk energy [27] for the sth window (using (4)):

$$\mathrm{JE}_s = J_{s1}^2 + J_{s2}^2 + J_{s3}^2, \tag{5}$$

Here, "jerk energy (JE)" is computed with 50% overlap and sliding window (4 seconds) based method. In addition, we compute 1st and 2nd derivative with respect to time for speed, acceleration (both longitudinal and lateral), jerk, jerk energy, and angular speed. From the above list, all redundant computations are removed. Thus, the primary measured values and the secondary computed values constitute the trip level data for further classification. All these data are denoted as $d \in D$; where D is the set of all data.



(a)

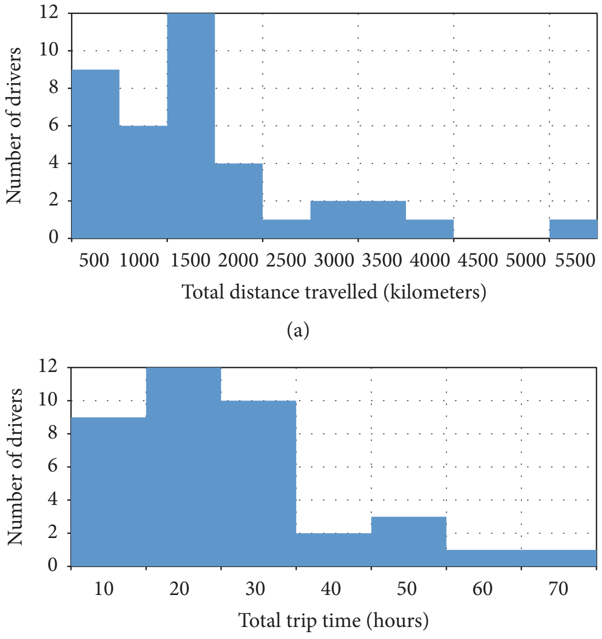
(b)

FIGURE 1: *Distribution of important parameters of journey for 2 drivers.* (a) Distribution of average speed. (b) Trips occurring at various time of day.

3. Method of Analysis

3.1. Basic Analysis and Feature Extraction. It is observed that the overall dataset is quite diverse, accommodating scenarios like driving on weekdays (or weekends), different time slots of the day, and different localities. Basic statistics derived from the driving data shows variation across drivers. Figure 1(a) shows the variation in the average speed (for each completed trip) for two drivers, namely, D017 and D035. For each driver, multiple trips are available and from every trip the average speed value is computed. Thus for each driver a set of average speeds (i.e., average speed per trip) is obtained. From this set, an empirical distribution of average speed can be computed. Also D017 and D035 have trips spread throughout different times of a day, as evident from Figure 1(b).

We have explored the dataset to obtain basic distributions of the total trip duration as well as the total distance covered by the participant drivers. This is shown in Figure 2. It can be seen that the typical distance covered by a driver is less than 1000 Km while typical driving time is approximately 20 hours. In order to statistically explore the dataset, we extracted multiple features from the driving data. We consider GPS measured "Speed" and "Heading" as primary data. From "Speed", we compute secondary data, namely, jerk, jerk energy, lateral acceleration, angular speed, and longitudinal acceleration. The positive and negative acceleration (both longitudinal and lateral) are segregated and separately treated. Next, the 1st and 2nd derivative (with respect to time) for all the above stated data are calculated; these also form secondary data. It is to



(a)

(b)

Figure 2: *Distribution of total journey distance and time for the participants.* (a) Histogram of total distance travelled by the drivers. (b) Histogram of total trip duration for the drivers.

be noted that, for "Heading," we take only the derivatives as secondary data. Subsequently the statistical features of the primary and secondary data, corresponding to every completed trip, are extracted. These features are mean, median, skewness, kurtosis, standard deviation, max, min, 97.5th percentile, 1st and 3rd quartiles (Q1, Q3), IQR (interquartile range: Q3–Q1), and 2.5th percentile. Table 2 summarizes the list of features. Thus, the variations in speed and acceleration are taken to be the major indicators of how the car is driven.

Some examples of features are median of speed, Q1 of 1st derivative of lateral acceleration, kurtosis of 2nd derivative of jerk energy, and 2.5th percentile of 2nd derivative of jerk energy. Our choice of suitability of the "feature" rests on the requirement that it should have good individual predictive ability. Also, the set of selected features should have low correlation amongst themselves and high correlation with class (driver in our case) [28]. As the next step, "analysis of variance" (ANOVA) is performed on the entire feature set in order to identify the statistically invariant features across all the drivers. For example, skewness of acceleration is seen to be very less close to 0 and does not vary with trip or driver. These are omitted. After such omission, we obtain a set of 137 features that is considered suitable for further analysis. From here onwards, these 137 features constitute a "set of all features" hitherto named as "global" feature set. It is to be noted that the proposed approach is dependent on the completion of the trip thereby excluding those applications where real-time identification of the driver is required. That is why dimensionality reduction methods like PCA are not considered. Also, the car type for each user is not known except for the fact that these are owned by the participant drivers. Hence, ANOVA cannot be performed with respect to car type. Throughout this study, we assume that the observed variations are due to differences in driving styles only. However, if the knowledge on the dynamics of the car as well as the road condition was available, it could have enhanced the quality of interpretation.

3.2. Classification for Driver Identification. Driver identification is approached as a classification problem, where each driver represents a class. From the trip data corresponding to a driver (belonging to a group of 4-5 drivers), the features are computed and classification is performed. Here, the classification accuracy with respect to a driver is measured by the percentage of correctly classified trips for that driver in a group. Similarly, the group level accuracy is measured by the percentage of correctly classified trips corresponding to that group. A driver misclassification reduces the overall accuracy.

To check the applicability of the global feature set, we performed Random Forest [29] based evaluation and observed that the accuracy obtained is good for the purpose of behavior identification. Throughout this study, for the purpose of classification, we have used only Random Forest with 100 trees and a batch size of 100. Maximum depth of the tree is set to unlimited. Table 3 gives the confusion matrix for a group of four drivers (a group consisting of drivers D001, D002, D006, and D009). Metrics like Precision, Recall, and *F*1 score are used for measuring performance of classifiers [30].

Overall accuracy obtained for this group is 77.7%, where accuracy is defined as the ratio of total number of correctly identified trips over the total number of trips. Within a group, the identification accuracy of individual driver is denoted by the computed value of *Recall*, percentage of his/her own trip identified correctly. Additionally, we compute *Precision* which is defined as how many trips (ratio), identified as belonging to a particular driver, actually belongs to him or her. For the purpose of multiclass classification, we computed the macro average (averaging the evaluation measures) performance measure [31]. Macro averages of Precision, Recall, and *F*1 score are 0.79, 0.76, and 0.78, respectively. We also observed that the other techniques like KNN (*k*-Nearest Neighbor) and SVM (Support Vector Machine) show lower accuracy as compared to Random Forest. {SVM is used with polynomial kernel and tolerance parameter is set to 0.001, iterations continue till convergence happens while KNN is used with number of nearest neighbors set to 8 and linear search} The models are then validated using *k*-fold cross-validation technique, where $k = 10$ is used. There is no algorithm that performs better in general; thus there is a need to choose an appropriate algorithm that outperforms others for the proposed problem definition. Random Forest outperformed SVM and KNN each time in our study. Similar observations have been identified by other authors [8, 16]. In [8], Hallac et al. mention that as the driver pool increases, the alternative approaches using Multinomial Logistic Regression and Support Vector Machines "dropped off significantly" as compared to Random Forest. In [16], Junior et al. compared Artificial Neural Networks, Support Vector Machines, Random Forest, and Bayesian Network towards driver behavior profiling. Junior et al. conclude

Table 2: Summary of features for driver identification.

Data	Feature
Speed, jerk, jerk energy, lateral acceleration, angular speed, and longitudinal acceleration	Mean, median, skewness, kurtosis, standard deviation, max, min, 97.5th percentile, 1st and 3rd quartiles (Q1, Q3), IQR (interquartile range: Q3–Q1), and 2.5th percentile
Positive lateral acceleration, negative longitudinal acceleration, positive lateral acceleration, and negative longitudinal acceleration	
All 1st and 2nd derivatives of data in rows 1 and 2	

Table 3: Accuracy and confusion matrix on a group (G4) of drivers using global feature set.

Driver ID	Predicted				Classification accuracy (Recall%)
	D001	D002	D006	D009	
Actual					
D001	31	1	7	0	79.5
D002	1	15	0	10	57.7
D006	11	0	37	0	77.1
D009	1	1	1	32	91.4
Precision (%)	70.5	88.2	82.2	76.2	

that Random Forest outperforms others for majority of assemblies.

In our application, the final accuracy is validated on "test data" (data from new journey of the drivers). It is found that the accuracy remains similar to what was obtained in k-fold cross-validation. Hence, for the rest of this paper, we present our evaluation using the global feature set together with Random Forest.

4. Results and Analysis

4.1. Driver Identification Results. For the purpose of driver identification, we segregated the drivers into different natural groups and analyzed the trips for each group. While grouping the drivers, we fundamentally looked at their regions of operation as well as the trip timings. Mostly, the drivers are grouped according to geographical similarity. They drive in the same area of a city with comparable trip timings; that is, the drivers operating in similar locality with similar departure time are grouped together. Figure 3 displays a map outlining the traversed regions for some of the groups. It needs to be mentioned that a few drivers fall in multiple groups as they have overlapping route locations and/or timings.

The blue marks in Figure 3 show the traversed routes by the participants. The red colored boundaries denote the zone covered by all the drivers belonging to the specific group. The underlying premise is that all the drivers in one group will face very similar external perturbations that may affect their driving styles. Therefore, the distinctiveness in their driving data can be attributed to their own propensities only.

In this section, we present the results obtained for each group of drivers as well as the accuracy obtained for individual drivers. These results are summarized in Table 4.

The detailed results for G4—the group consisting of drivers D001, D002, D006, and D009—have been presented in Table 3. Across all the groups, G1 to G9, we obtain an average overall accuracy of 82.3%. This in turn confirms the strength of the selected feature set as well as the classification method chosen for the purpose of identification. At group level, the least average accuracy obtained is 74%, being that for group 9. The average accuracy in group 9 is affected by one driver, D055. For driver D055, none of his trips are classified correctly (total 9 trips are considered). This happened due to the limited availability of training data. Although the group level average accuracy is deemed to be high (>78% except for G9); the same cannot be unequivocally stated for individual driver identification. Within a group, both Recall and Precision are observed to vary, with the minimum value being 0.25 (for G9). However, for all such cases, the available training data is much smaller than others.

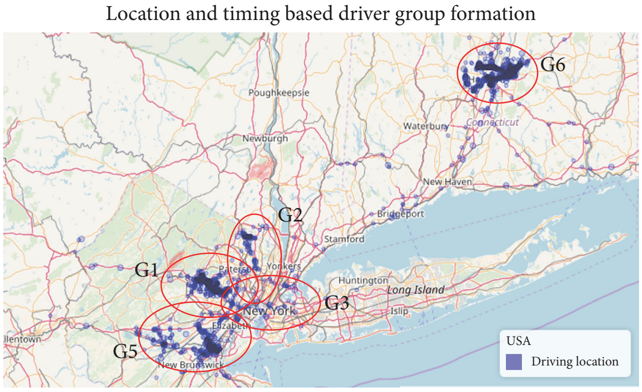


Figure 3: Map displaying geographical locations for the group of drivers [typical journey localities for each group of drivers are circled].

It is seen that the proposed method works quite well when a group of limited size is formed. Cross-validation results for a sample group are shown in Tables 5 and 6 for 10-fold cross-validation and tested on new data.

We have placed D021 in three groups (G1, G3, and G5) while D014 is placed in two groups (G1 and G2). This is because the routes traversed by the drivers in these groups lie in reasonable proximity and the routes undertaken by D021

Table 4: Summary of driver identification scores for all groups.

Group	Precision	Recall	*F*1 score	True negative rate	Overall accuracy
G1	0.81	0.79	0.8	0.93	0.82
G2	0.86	0.74	0.79	0.95	0.84
G3	0.80	0.79	0.79	0.93	0.8
G4	0.79	0.76	0.77	0.94	0.78
G5	0.81	0.78	0.79	0.94	0.79
G6	0.86	0.81	0.83	0.93	0.83
G7	0.86	0.90	0.88	0.97	0.9
G8	0.9	0.89	0.89	0.97	0.91
G9	0.45	0.63	0.87	0.89	0.74

Table 5: Accuracy and confusion matrix on group 6 (G6) of drivers with all features.

Driver ID	Predicted				Recall (%)
	D0068	D0069	D0064	D0066	
Actual					
D0068	38	2	1	0	92.6
D0069	4	22	1	0	81.4
D0064	3	4	13	0	65
D0066	1	1	0	13	86.6
Precision (%)	82.6	75.8	86.6	100	

Table 6: Accuracy and confusion matrix on group 6 (G6) of drivers for 10-fold cross-validation.

Driver ID	Predicted				Recall (%)
	D0068	D0069	D0064	D0066	
Actual					
D0068	42	0	22	8	58.3
D0069	0	42	2	0	95.4
D0064	8	0	118	18	81.9
D0066	3	0	25	70	71.4
Precision (%)	79.2	100	70.6	72.9	

and D014 fall in any one of these overlapping groups. However, for the purpose of identification, we have considered all the trips undertaken by the said drivers (like 14 trips for D021 and 11 trips for D014) in each of their overlapping groups; else a group wise route segregation would have resulted in a substantially reduced data for the given drivers in these groups.

It is also of interest to evaluate the identification accuracy when all the 38 drivers are placed in a single group. Towards this objective, we apply Random Forest on the entire dataset. It is seen that 25 drivers displayed accuracy better than 40% with 16 of them crossing 50% accuracy level. The median accuracy obtained is 0.46 with standard deviation being 0.22. Figure 4 represents the confusion matrix in terms of a heatmap which shows how many (percentage) trips of a driver are correctly classified and how many of them are mapped into other drivers.

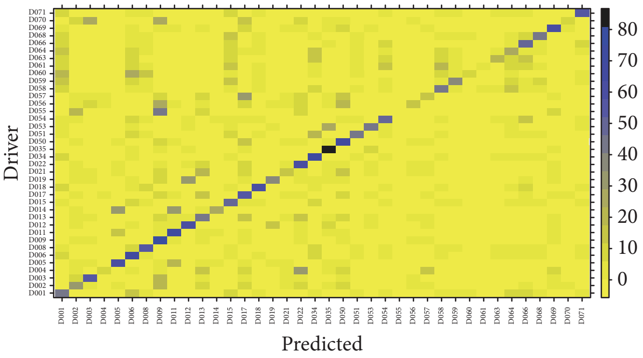


Figure 4: Accuracy and classification obtained for all 38 drivers taken together in a single group.

4.2. Effect of Skill and Aggression on Driver Identification Accuracy. In our previous work [19], we have quantified each driver in terms of 2D factor analysis—Aggression and Skill. High acceleration/deceleration maneuvers induce risk in driving. These patterns get reflected as heavy tail of the acceleration distribution obtained from trip level data. Since kurtosis is a good measure of sharp peak and heavy tail, it is chosen as a suitable feature to be extracted from the acceleration profile. Also, the combination of kurtosis and skewness squared can be used as an identifier of underlying distribution. For a skilled driver, the basic nature of acceleration will not significantly vary between trips. In our study, it is found that skewness of acceleration profile is close to 0 (pdf of acceleration is symmetric). Therefore, only the variation in kurtosis over all trips of a driver indicates variation in underlying distribution of acceleration. The inverse of variation in kurtosis is quantified as a measure of skill and the median value of kurtosis is used as an indicator of aggression [19]. Similarly, we consider the median and standard deviation corresponding to lateral acceleration. Then for a group of drivers, these measures are normalized to obtain skill and aggression score. Finally, we form five clusters of drivers with 50% (19) of participants treated as "normal" drivers. From this analysis, we get 7 drivers in the cluster denoted as "high skill and low aggression" and 8 drivers in "low skill and high aggression" cluster. It is be noted that such clustering is only relative; if the population changes, some drivers may relocate to adjacent clusters.

Each driver, based on all the available trips, is assigned a point in two-dimensional space as aggression and skill. We now investigate whether relative skill and/or aggression can affect identification accuracy. It is to be noted that the skill-aggression scores for each driver are derived from the same GPS data which is used for identification. We have taken a snapshot of drivers for analysis. In Table 7, we display skill-aggression score [19] for all 38 drivers. Please note that Table 7 is sorted according to skill score, from highest relative skill to lowest, with the scores being bounded to ±4: skill score of +4 is termed as most skillful (or most aggressive) and

Table 7: Aggression and skill score for drivers.

Driver	Aggression	Skill	Associated group
D018	−1.06	3.87	3
D053	−1.62	1.76	5
D066	−0.98	1.65	6
D071	−1.6	1.58	9
D058	−0.41	1.39	1
D019	0.46	0.9	3
D008	−1.06	0.76	7, 8
D054	−0.53	0.41	7, 8
D068	−0.85	0.35	6
D035	−0.94	0.31	Ungrouped
D012	−0.09	0.23	2
D055	0.33	0.21	9
D006	−0.61	0.1	4
D015	−0.4	0.02	Ungrouped
D059	−1.11	−0.05	8
D063	−0.49	−0.15	9
D064	−0.43	−0.19	6
D069	−0.31	−0.2	6
D003	−0.24	−0.22	Ungrouped
D060	−0.25	−0.28	Ungrouped
D009	−0.05	−0.32	4
D001	−0.82	−0.37	4
D013	0.35	−0.44	2
D057	1.25	−0.47	Ungrouped
D002	0.73	−0.52	4
D034	0.64	−0.55	1
D017	0.61	−0.59	3
D070	3.18	−0.6	7, 9
D056	0.87	−0.69	5
D061	−0.91	−0.69	8
D011	0.46	−0.77	Ungrouped
D022	0.83	−0.82	2
D004	0.92	−0.84	7
D005	1.4	−0.85	Ungrouped
D021	0.22	−0.94	1, 3, 5
D014	2.01	−0.97	1, 2
D050	0.64	−0.97	5
D051	−0.09	−1.03	5

Table 8: Confusion matrix on group 1 (G1) of drivers.

Driver ID	Predicted				Recall (%)	Skill rank
	D014	D021	D034	*D058*		
Actual						
D014	6	5	0	0	54.4	4
D021	1	13	0	0	92.8	3
D034	0	1	27	3	87.1	2
D058	0	1	5	26	*81.2*	*1*
Precision (%)	85.7	65	84.3	*89.6*		

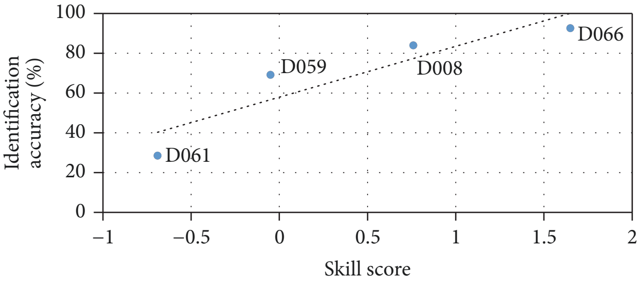


Figure 5: Identification accuracy obtained in a group with identical aggression displayed by the drivers.

−4 depicting the least skill (or minimum aggression). These scores are so normalized that an average driver will get a score of (0, 0).

We take the case of four drivers, namely, D058, D018, D066, and D071, who display high relative skills and moderate (to low) aggression. In the present case, these drivers got grouped into G1 (Table 8), G3 (Table 9), G6 (Table 10), and G9 (Table 11), respectively.

From the results presented in Tables 8 to 11, we observe that the identification accuracy *(% Recall)* for each of these high skilled drivers exceeds 80%, which is deemed very impressive. Does it therefore mean that certain behavioral attributes (like higher skill and low aggression) ensure significantly better identification in a group? Let us examine this further by forming arbitrary groups, that is, groups formed solely based on their relative standing in terms of aggression and skill.

We investigated the above hypothesis by forming a group of four drivers (D066, D008, D0059, and D061) who display similar aggression levels but widely separated in skill scores (refer to Table 7). Figure 5 shows the results. It is seen that, for such a group, the driver with highest skill factor can be identified as accurately as 93%. There is also a monotonic increase in identification accuracy with respect to increasing skill factor.

In order to test the hypothesis further, we conducted an investigation by forming a group with the four most skillful drivers, namely, D018, D071, D066, and D053. They are also closely spaced in 2D-aggression/skill plane, meaning thereby that the key attributes of their driving styles are also similar. The result is shown in Figure 6. The achieved accuracy for each of them is very high (>80%)

Considering these results in conjunction with the ones obtained for the same drivers in their natural groupings, one may tend to infer that the drivers displaying very high skill factor (along with moderate aggression) are accurately identifiable, irrespective of the group composition of similar size. However the effect of larger group size on the stated hypothesis needs further study.

In a similar vein, we investigate a group of drivers displaying lowest skills as well as being closely spaced in the 2D-aggression/skill plane. These are D004, D022, D050, and D056. The result is shown in Figure 7.

Comparing the results obtained from Figure 7 with their natural groups, it is found that, for the drivers with lower

TABLE 9: Confusion matrix on group 3 (G3) of drivers.

Driver ID	Predicted				Recall (%)	Skill rank
	D017	*D018*	D019	D021		
Actual						
D017	21	0	0	3	87.5	3
D018	0	13	0	0	*100*	*1*
D019	1	2	9	2	64.2	2
D021	4	0	1	9	64.2	4
Precision (%)	80.7	*86.6*	90	64.2		

TABLE 10: Confusion matrix on group 6 (G6) of drivers.

Driver ID	Predicted				Recall (%)	Skill rank
	D064	*D066*	D068	D069		
Actual						
D064	38	2	1	0	92.6	3
D066	4	22	1	0	*81.4*	*1*
D068	3	4	13	0	65	2
D069	1	1	0	13	86.6	4
Precision (%)	82.6	*75.8*	86.6	100		

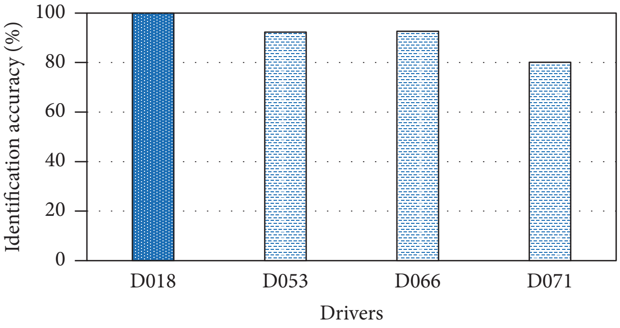


FIGURE 6: Identification accuracy obtained in a group of drivers with very high skills.

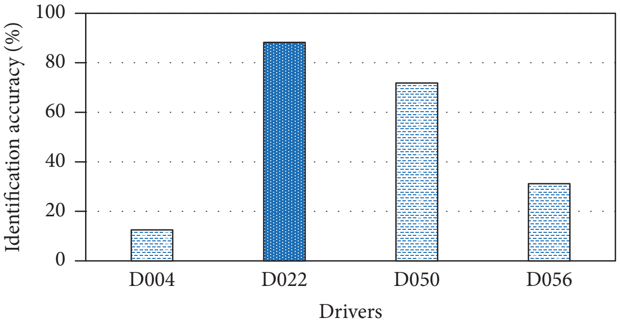


FIGURE 7: Identification accuracy obtained in a group of drivers with low skills and similar aggression.

skills, one may not be able to predict a given range of accuracy measures. These drivers are inconsistent in their driving styles. For all such cases, group composition does matter. Therefore we can state that the skill and aggression factors as seen in driving styles have considerable effect on driver identification accuracy.

4.3. Discussion. The objective of the presented work is to investigate whether we are able to uniquely identify a driver (amongst a group of drivers) solely based on the GPS data covering the entire trip. Our study shows that, even with such minimal sensing, it is possible to identify a driver with an accuracy of nearly 78%. The proposed method is not envisioned as a replacement for biometric authentication. The purpose of this study is to accurately identify a driver based on the statistical features of GPS data alone in contrast with the multisensor approach taken by previous authors. It is seen that, for a small group of drivers, an automated driver identification routine is feasible even with access to GPS data alone. At the same time, the proposed method throws some challenges. For example, since trip based feature set is only considered, its applicability to real-time monitoring scenarios remains doubtful.

We now compare our results with relevant prior works. Driver identification [21] by Burton et al. achieves 95% confidence interval on 10 drivers using a simulated data set. Fung et al. [22] achieve 61% accuracy for older (70+ years) drivers using only acceleration and speed. Moreira-Matias and Farah [23] achieve 88% accuracy on a large multisensory data obtained from 217 drivers. Moreira-Matias and Farah [23] also develop a generic machine learning methodology and feature reduction process. Martínez et al. [24] give accuracy above 80% with multisensor data on 11 drivers. A comparative analysis of our work with others is given in Table 12.

From Table 12, it is seen that the accuracy attained through our method is quite comparable to the published works. While the results are very promising, it is felt that additional dedicated sensors like brake pedal and steering

Table 11: Confusion matrix on group 9 (G9) of drivers.

Driver ID	Predicted				Recall (%)	Skill rank
	D055	D063	D070	*D071*		
Actual						
D055	0	0	2	0	0	2
D063	0	8	1	3	66.6	3
D070	0	0	1	0	100	4
D071	0	4	0	20	*83.3*	*1*
Precision (%)	NA	66.6	25	*86.9*		

Table 12: Comparison of the proposed method with other existing methods.

Method	Sensors	Data	Obtained accuracy
Our method	GPS only	38 drivers, 2 months	82.3%
Enev et al. [6]	Brake pedal	15 drivers	87%
Zhang et al. [7]	Multisensory	20 drivers, simulated driving	85%
Hallac et al. [8]	12 sensor readings, other than GPS	10 cars, 64 drivers	76% (for 2 drivers) 50% (for 5 drivers)
Burton et al. [21]	Multisensor, brake pedal, gas pedal, speed, etc.	10 drivers, simulated driving	95% confidence interval
Fung et al. [22]	Location and speed	14-old-age drivers	30–61%
Moreira-Matias and Farah [23]	Multisensor data	Dataset of 217 families collected over one year	88%
Martínez et al. [24]	Multisensor data	11 different drivers 25 Km (40 min.) data	Above 80%

wheel will certainly increase the identification accuracy to a large extent. Our data set does not contain any information about the traffic and weather conditions, when the trips were undertaken. Such additional information could have been utilized for natural group formation possibly leading to superior results.

5. Conclusions

In this paper, we present the results of our study on driver identification approach using only smartphone's GPS. Data collected from 38 drivers for a duration of two months and covering a total journey length of 50000 Km have been analyzed. This study demonstrates that a set of 137 features, extracted from the GPS data corresponding to the completed trips, can be used to identify the driver of a vehicle. For the purpose of driver identification, we have segregated the drivers into natural groups of four to five drivers in each group where route proximity is used as the deciding factor for such segregation. The investigations have offered important insights regarding driving behavior and its role vis-à-vis our ability to identify the driver accurately. We have observed that certain behavioral attributes (like higher skill, low aggression) ensure significantly better identification accuracy. The analysis presented in this paper shows that, for the "driver identification" problem, an average accuracy of 82.3% is achieved, where the drivers are grouped into multiple groups (each group has 4-5 drivers). Thus using only GPS, a good accuracy is obtainable and GPS can serve as a backbone of driver identification system. Further improvements can be achieved by using other sensors with GPS.

We conclude by stating that it is possible to identify drivers with reasonably high accuracy even when only the smartphone GPS is used. At the same time, it is to be noted that much more investigations are needed in the future in order to identify a driver with near certainty. It is to be seen whether the additional sensory information available in a vehicle's ECU (Electronic Control Unit) can be fused with this GPS data so that this accuracy is greatly enhanced.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] D. J. Watts, "Should social science be more solution-oriented?" *Nature Human Behaviour*, vol. 1, no. 1, p. 0015, 2017.

[2] J. Wahlström, I. Skog, and P. Handel, "Smartphone-Based Vehicle Telematics: A Ten-Year Anniversary," *IEEE Transactions on Intelligent Transportation Systems*, 2017.

[3] R. Fuller, "Towards a general theory of driver behaviour," *Accident Analysis & Prevention*, vol. 37, no. 3, pp. 461–472, 2005.

[4] J. Fuller, "Psychology and the highway engineer," *Human Factors for Highway Engineers*, pp. 1–10, 2002.

[5] N. Lin, C. Zong, M. Tomizuka, P. Song, Z. Zhang, and G. Li, "An overview on study of identification of driver behavior characteristics for automotive control," *Mathematical Problems in Engineering*, vol. 2014, Article ID 569109, 2014.

[6] M. Enev, A. Takakuwa, K. Koscher, and T. Kohno, "Automobile Driver Fingerprinting," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 1, 2016.

[7] X. Zhang, X. Zhao, and J. Rong, "A study of individual characteristics of driving behavior based on hidden markov model," *Sensors & Transducers Journal*, vol. 167, no. 3, pp. 194–202, 2014.

[8] D. Hallac, A. Sharang, R. Stahlmann et al., "Driver identification using automobile sensor data from a single turn," in *Proceedings of the 19th IEEE International Conference on Intelligent Transportation Systems, ITSC 2016*, pp. 953–958, Brazil, November 2016.

[9] J. Engelbrecht, M. J. Booysen, G.-J. Van Rooyen, and F. J. Bruwer, "Survey of smartphone-based sensing in vehicles for intelligent transportation system applications," *IET Intelligent Transport Systems*, vol. 9, no. 10, pp. 924–935, 2015.

[10] P. Themann, J. Bock, and L. Eckstein, "Optimisation of energy efficiency based on average driving behaviour and driver's preferences for automated driving," *IET Intelligent Transport Systems*, vol. 9, no. 1, pp. 50–58, 2015.

[11] N. Abuali and H. Abou-Zeid, "Driver behavior modeling: Developments and future directions," *International Journal of Vehicular Technology*, vol. 2016, Article ID 6952791, 2016.

[12] S. Laapotti, E. Keskinen, M. Hatakka et al., "Driving circumstances and accidents among novice drivers," *Traffic Injury Prevention*, vol. 7, no. 3, pp. 232–237, 2006.

[13] E. I. Vlahogianni and E. N. Barmpounakis, "Driving analytics using smartphones: Algorithms, comparisons and challenges," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 196–206, 2017.

[14] T. Tanprasert, C. Saiprasert, and S. Thajchayapong, "Combining Unsupervised Anomaly Detection and Neural Networks for Driver Identification," *Journal of Advanced Transportation*, vol. 2017, pp. 1–13, 2017.

[15] P. Phumphuang, P. Wuttidittachotti, and C. Saiprasert, "Driver identification using variance of the acceleration data," in *Proceedings of the 19th International Computer Science and Engineering Conference, ICSEC 2015*, Thailand, November 2015.

[16] J. F. Junior, E. Carvalho, B. V. Ferreira et al., "Driver behavior profiling: an investigation with different smartphone sensors and machine learning," *PLoS ONE*, vol. 12, no. 4, Article ID e0174959, 2017.

[17] A. Chowdhury, T. Chakravarty, and P. Balamuralidhar, "Estimating true speed of moving vehicle using smartphone-based GPS measurement," in *Proceedings of the 2014 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2014*, pp. 3348–3353, usa, October 2014.

[18] A. Chowdhury, A. Ghose, T. Chakravarty, and P. Balamuralidhar, "An improved fusion algorithm for estimating speed from smartphone's Ins/Gps sensors," *Next Generation Sensors and Systems*, vol. 16, pp. 235–256, 2015.

[19] T. Banerjee, A. Chowdhury, and T. Chakravarty, "Mydrive: Drive behavior analytics method and platform," in *Proceedings of the 3rd International Workshop on Physical Analytics, WPA 2016*, pp. 7–12, Singapore.

[20] R. Vaiana, T. Iuele, V. Gallelli, and D. Rogano, "Demanded versus assumed friction along horizontal curves: An on-the-road experimental investigation," *Journal of Transportation Safety & Security*, pp. 1–27, 2017.

[21] A. Burton, T. Parikh, S. Mascarenhas et al., "Driver identification and authentication with active behavior modeling," in *Proceedings of the 12th International Conference on Network and Service Management, CNSM 2016 and Workshops, 3rd International Workshop on Management of SDN and NFV, ManSDN/NFV 2016 and International Workshop on Green ICT and Smart Networking, GISN 2016*, pp. 388–393, Canada, November 2016.

[22] N. C. Fung, B. Wallace, A. D. C. Chan et al., "Driver identification using vehicle acceleration and deceleration events from naturalistic driving of older drivers," in *Proceedings of the 12th IEEE International Symposium on Medical Measurements and Applications, MeMeA 2017*, pp. 33–38, USA, May 2017.

[23] L. Moreira-Matias and H. Farah, "On Developing a Driver Identification Methodology Using In-Vehicle Data Recorders," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 9, pp. 2387–2396, 2017.

[24] M. V. Martínez, J. Echanobe, and I. Del Campo, "Driver identification and impostor detection based on driving behavior signals," in *Proceedings of the 19th IEEE International Conference on Intelligent Transportation Systems, ITSC 2016*, pp. 372–378, Brazil, November 2016.

[25] P. Handel, I. Skog, J. Wahlstrom et al., "Insurance telematics: Opportunities and challenges with the smartphone solution," *IEEE Intelligent Transportation Systems Magazine*, vol. 6, no. 4, pp. 57–70, 2014.

[26] R. Vaiana, T. Iuele, V. Astarita et al., "Driving behavior and traffic safety: an acceleration-based safety evaluation procedure for smartphones," *Modern Applied Science (MAS)*, vol. 8, no. 1, pp. 88–96, 2014.

[27] T. Chakravarty, A. Chowdhury, A. Ghose, C. Bhaumik, and P. Balamuralidhar, "Statistical analysis of road-vehicle-driver interaction as an enabler to designing behavioral models," *International Journal of Modeling, Simulation, and Scientific Computing*, no. 2, 2014.

[28] A. Hall Mark and L. A. Smith, "Practical feature subset selection for machine learning," in *Computer Science '98 Proceedings of the 21st Australasian Computer Science Conference ACSC'98, Perth*, C. McDonald, Ed., pp. 181–191, Springer, Berlin, 1998.

[29] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[30] D. M. W. Powers, "Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[31] V. Van Asch, "Macro-and micro-averaged evaluation measures," Tech. Rep., 2013, http://www.cnts.ua.ac.be/~vincent/pdf/microaverage.pdf.