Driver authentication by quantifying driving style using GPS only

Tanushree Banerjee, Arijit Chowdhury, Tapas Chakravarty, Avik Ghose
TCS Research & Innovation, Kolkata, India
Email: (tanushree.banerjee, arijit.chowdhury2, tapas.chakravarty, avik.ghose)@tcs.com

***Abstract*—Driver authentication that is verifying a driver's identity constitutes an important aspect of modern day automobile. A driver gets access to drive a car based on his identity. Identity is normally verified with smart card or possession of key. Also there exist identity verification based on fingerprint, passcode and image based approach using camera mounted inside a car. However such approach do not consider driving style based authentication. In this paper, we present an approach that uses personalized statistical feature set extracted from the global positioning system (GPS) data to authenticate a driver. Such personalized feature set reduces computation, improves interpretability of features and accuracy. Proposed method is further enriched by determining and using the most suitable machine learning technique. Our approach is tuned to increase accuracy, sensitivity and specificity. Overall mean area under receiver operating characteristic curve (AUC) obtained is 0.9 which implies the robustness of this technique. Primary contribution of the paper is personalized feature set for driver authentication. We provide performance comparison of several machine learning algorithms such as SVM, Random Forest, Naïve Bayes, MLP etc. for driver authentication.**

***Index Terms*—Global Positioning System (GPS), driving style, authentication, machine learning**

I. Introduction

Driving is one of the most common human activity. Identity verification of the driver is important in order to prevent theft cases. Knowledge of drivers identity can also help in designing customized instructions for better operation of the vehicle. There has been research works on facial recognition system to prevent auto theft and driving fraud [1], [2]. Driver identification system based on telematics data is an interesting field of study. Vehicle telematics mostly uses dedicated sensors. Modern smartphones are designed with inbuilt sensors, which leads to many interesting applications in smartphone based vehicle telematics. Wahlström [3] gave a detailed review of smartphone based vehicle telematics. Enev [4] used multi sensor data to identify drivers with an accuracy of 87% (99% with top 5 sensors). Zhang [5] achieved 85% accuracy using hidden markov model (HMM). Identification of a driver in a group of size 4-6 was attempted in [6] with average accuracy of 82.3%. Recently authors in [7], have implemented classification and feature selection for driver identification.

These research work focused on finding 'who the driver is' among a group of drivers. In recent times, driver fraud i.e. driving a car using a false identity of another driver is increasing. That calls for the requirement of authentication of the driver instead of identification. Authentication is slightly a different problem than identification. The objective of driver authentication is to generate an alarm if a driver profile does not match with the profile unique to the original driver. There has been cases of vehicle theft, lowering insurance premium through driver impersonation with fake identity. In order to get rid of such cases, we propose a behavioral biometric which authenticates a driver based on his/her natural driving style. This privacy preserving method has great implications for detecting driver fraud problem and would prevent such impersonations and corruptions in the vehicle sector.

The core objective of this research is to authenticate a driver with minimal infrastructure/deployment requirement. So only GPS data is used as input. That is, given a trip level GPS data, we attempt to respond to the query *'whether driver X is driving the car?'*. Towards that, our investigations lead us to identify a unique (or personalized) feature set for each driver, based on his driving style, and then later use them to authenticate the driver. We complete our study by recommending the particular statistical features that are important for authentication of driver. Proposed scheme can be easily implemented on large scale and would definitely lead the way for minimal sensing approach in vehicle telematics field.

Some advantages of proposed method for driver authentication are:

1) Privacy preservation: No credentials or image of the driver is stored.
2) Minimal infrastructure usage: Only GPS is required.
3) Easy to deploy: No active participation is required from the user end.

In this paper, driver authentication is treated as a two class classification problem (i.e. yes and no) for each driver. Main contribution of the authors is authenticating a driver, based on GPS data from a trip driven by him/her only. Several machine learning techniques are used to analyze the features extracted from the GPS data. Random Forest based classification provided the best results. Remainder of the paper is structured as follows. In section II data collection, methodology used and extracted features are listed briefly. In section III, proposed method is presented and evaluated. Robustness of the proposed model is checked and confirmed after performing multiple tests. Conclusion and future directions of our work are presented in section IV.

II. Data and basic analysis

A. Experimental setup and data collection

38 drivers participated in this study, and only GPS data (timestamp, speed, latitude, longitude and heading) for each trip is considered. Total data consists of around 4000 trips (more than 50000Kms) collected in USA from associates. No personal data such as name, age etc. were made available to us. Experimental setup and dataset analyzed is the same as used in [6]. Authors had no involvement with data collection process. Most of the trips are taken on weekdays for over a period of 2 months i.e. around 9 weeks of data is collected. In this paper, briefly some aspects of the data is mentioned. For a detailed description, please refer to [6].

GPS data such as course, horizontal accuracy, timestamp, altitude, latitude, longitude, and speed in m/s are logged by the data collection unit(i.e. GPS logger) at 1 Hz rate. Collected data is annotated using a driver ID (unique ID for each driver). Then from this primary data, secondary data is computed, which consists of longitudinal and lateral acceleration, angular speed, jerk and jerk energy [6], [8] and their 1st and 2nd derivatives. Trip level data constitutes primary (measured values) and the secondary (computed from the measured values) corresponding to a given trip. After this step, each trip is quantified to a set of features, which are used to authenticate a driver.

B. Feature Extraction

Statistical exploration of the data is done, by extracting multiple features from the data. In total 137 features are extracted for each driver, that are related to their unique driving style. This global feature set is denoted by $\mathrm{FS_{Global}}$. Each of these feature is important for driver authentication problem, but all features are not important for every driver. Therefore, in present work we focus on identifying features that more related to a particular driver and thus can be used to authenticate him/her. For each driver with driver ID D_i (for i=1,2...,38) we construct a personalized feature set $\mathrm{FS_{PersonalD_i}}$ which is a subset of $\mathrm{FS_{Global}}$.

Proposed method for authentication is validated by k-fold cross validation (with k=10). Robustness is confirmed by analyzing sensitivity, specificity and receiver operating characteristic (ROC) analysis. Results are provided in section III. Amongst 9 weeks of data, initial 7 weeks is used for analysis and classifier selection, followed by methodology validation. Remaining 2 weeks of data are kept separately for further validation of final model (i.e. personalized features and classifier). Final model is validated on last 2 weeks dataset. Entire feature computation can be done in cloud or in a mobile device itself.

III. METHOD AND RESULTS

In this section we deal with the issue of authentically verifying a drivers identity. Let a particular trip is claimed to be driven by driver A. To verify the authenticity of the claim we try to answer the following question:

- Is the trip driven by driver A or not?

To answer this question, authors used a data set, consisting (initial 7 weeks data) having equal proportion of trips driven by A as well as by others. The model is validated using 10 fold cross validation. It is to be noted that this is not a biometric identification system. In real life scenario, it is highly probable that occasional impersonation (where the identities of a good and bad drivers are switched) takes place so as to improve driving score and thereby reduce insurance premium. Thus for the insurance providers, it is an important question whether the reported driver is driving or not.

Initially, it is hypothesized that driver specific features (for every driver) can lead to better results and that personalized features can be used to define his/her natural driving style. Also those features can serve as a distinguished style for that driver. Towards that goal, we tried to create a personalized feature set for every driver. For feature ranking purpose, R statistical software [9] is used, on the training data. To rank the features, Boruta package [10] is used. Boruta gives variable importance measure (VIM) for each of the features - corresponding to the given driver and also accepts (or rejects) a feature for a classification problem. Using Boruta, we are able to select the most significant features for each driver. It is then found that every driver has different set of significant features; here onwards referred to as personalized feature set ($\mathrm{FS_{PersonalDi}}$). For the next phase of our investigation, we used only the personalized feature set for each driver. These features define parameters of importance for respective driver which distinguishes him/her from the rest.

For each driver, cardinality of personalized feature set varies. A maximum of 56 features is selected for driver D006 whereas for D004 only 11 features are selected. The average number of selected feature is 30 with a standard deviation of 11. Table I gives a set of top 10 features for four drivers, representing the idea. We observe that four of them show distinct patterns in terms of the signature seen in their maneuvers. Following abbreviations have been used in Table I:

PosLatA $\equiv$ Positive Latudinal acceleration
PosLonA $\equiv$ Positive Longitudinal acceleration
NegLatA $\equiv$ Negative Latudinal acceleration
NegLonA $\equiv$ Negative Longitudinal acceleration
Jerk $\equiv$ First order derivative of acceleration with respect to time [6].
Pctl97.7 $\equiv$ 97.7th percentile
Diff $\equiv$ Differentiation with respect time
Diff2 $\equiv$ 2nd order differentiation with respect time

Derivative of course clearly relates to the steering wheel operations and negative longitudinal acceleration reflects a drivers style of using brake pedal. It is observed that D003 shows a propensity for lateral acceleration, D006, D009 and D011 display propensity for brake and steering wheel angle respectively; as apparent from their most important features. Fig. 1 further illustrates this fact, which shows variation of the particular feature, IQR of negative longitudinal acceleration across 19 drivers. Its clear that for drivers (say) D002, D004

TABLE I
TOP 10 SELECTED FEATURES FOR FEW DRIVERS

Feature Rank	D003	D006	D009	D011
1	Q2(PosLatA) [1]	Q3(NegLonA)	IQR(Range(Diff(Course)))	Pctl97.7(NegLonA)
2	Q2(Jerk(LatA))	Q1(NegLonA)	Q3(Diff(LatA))	Max(NegLonA)
3	Q1(JerkLatA)	Pctl97.7(NegLonA)	IQR(Diff(LatA))	Q3(NegLonA)
4	IQR(LatA)	Pctl97.7(PosLatA)	Q1(Jerk(LatA))	IQR(Diff(LonA))
5	Q3(PosLatA)	Q2(NegLonA)	Max(Speed)	Q1(Diff(LonA))
6	Q1(Diff2(Course))	Q3(Diff(LatA))	Q1(Diff(LatA))	Q3(Diff(LonA))
7	Q3(Diff(Course))	Pctl97.7(Jerk(LatA))	Q1(Diff(Course))	IQR(LonA)
8	Pctl97.7(Speed)	IQR(NegLonA)	Pctl97.7(NegLonA)	Q2(NegLonA)
9	IQR(PosLatA)	IQR(JerkLonA)	Max(NegLonA)	Q3(LonA)
10	Q1(Diff(LatA)	IQR(Diff(LatA))	IQR(Diff(Course))	Q1(NegLonA)

[1] Pctl97.7 = 97.7th percentile, Q1, Q2(Median) and Q3 = the 25th, 50th and 75th quartiles, IQR = Interquartile Range

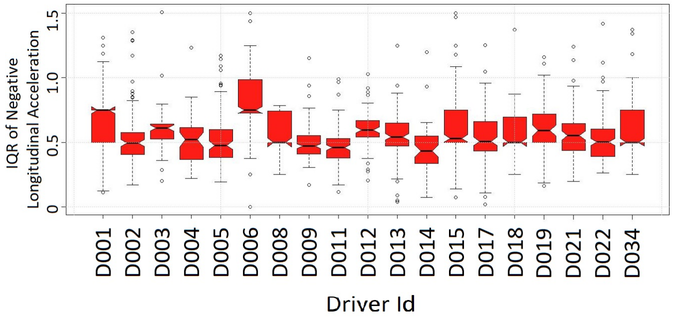

Fig. 1. Variation of IQR of negative longitudinal acceleration across selected 19 drivers

and D005, this particular feature value is almost equal, however for driver D006, the value is very different compared to his peers, which thereby gets reflected in Table I with a rank of 8. This implies each individuals propsensity to use brake pedal (which influences the negative longitudinal acceleration) is different, irrespective of the road being driven. Our deductions seems to tally with earlier findings by Enev [4]; where they identified the top sensors for unique driver identification as brake pedal, maximum engine torque, steering wheel and lateral acceleration, in a descending order of significance.

We try to authenticate a driver by using different classifiers. We had 38 drivers' trip data. For each driver, the classifier is trained on initial 7 weeks data, followed by 10 fold cross validation for different classifiers. Classifiers are used to identify if the same driver is driving or not; corresponding to a given trip. In the present analysis, sensitivity and specificity [11] for each driver is also measured along with overall accuracy. Sensitivity and specificity are defined as follows:

- Sensitivity: Proportion of a drivers true journey that are correctly classified.
- Specificity: Proportions of negatives (i.e. journey not driven by a particular driver) is correctly identified as false (i.e. not driven by that driver).

The following steps are used for selection of classifier:

1) Perform analysis on initial 7 weeks of the data, referred to as past dataset, amongst the 9 weeks data.
2) Compare accuracy, sensitivity and specificity, obtained from the analysis in order to choose the best classifier.
3) Then validate robustness of the chosen classifier by analysing the area under receiver operating characteristic curve analysis [12])

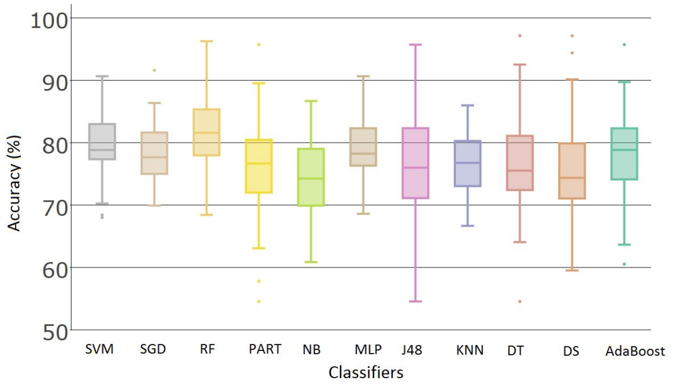

Fig. 2. Box Plot comparison for different statistical classifier on the dataset for 10 Fold Cross Validation. Y Axis denotes accuracy obtained in percentage. X Axis denotes different classifiers.

For every driver, model creation process includes validation of model with k fold cross validation on training data. We finally choose a classifier which maximized accuracy, sensitivity and specificity for 10 fold cross validation. For each driver, we obtain a separate unique model per classifier. In our work we selected 11 different classifiers to classify a journey driven by a driver or not. The list of classifiers investigated is as follows:

SVM (Support Vector Machine with Sequential minimal optimization), (RF)Random Forest [13], SGD (Stochastic Gradient Descent) [14], PART [15], MLP (Multilayer Perceptron) [16], J48 [17], KNN (i.e. IBK: k-Nearest Neighbors algorithm) [18], (DT)Decision Table [19], (DS)Decision Stump(i.e. a machine learning model consisting of a one-level decision tree), Naïve Bayes [20] and AdaBoost [21]. For each classifier, the classification accuracy is checked on 10 fold cross validation for all the drivers. Thus, 38 accuracy values are obtained per

TABLE II
ACCURACY FOR DIFFERENT CLASSIFIERS WITH DRIVER SPECIFIC FEATURE SET

Classifier	Median Accuracy	Standard Deviation
SVM	78.82	6.5_{Minimum}
SGD	77.66	8.7
Random Forest	83.01^{Maximum}	12.0
PART	76.67	8.2
Naïve Bayes	74.24	6.7
MLP	78.24	7.3
J48	75.99	8.5
IBK (kNN)	76.77	8.7
Decision Table	75.52	13.1
Decision Stump	74.37	8.9
AdaBoost	78.84	8.1

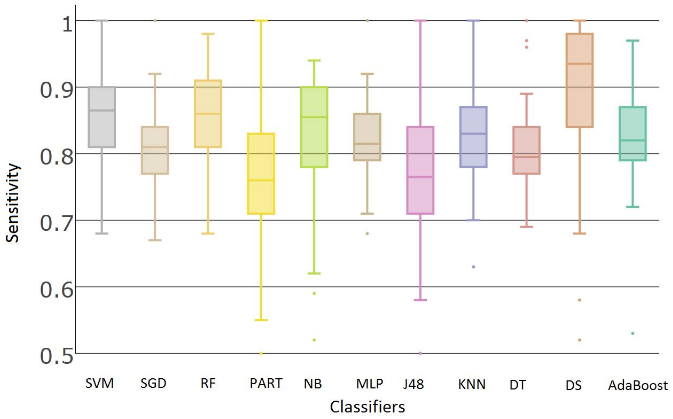

Fig. 3. Sensitivity value using different classifiers obtained for 38 Drivers. Y Axis shows the value of obtained sensitivity in percentage and X Axis shows the algorithm used

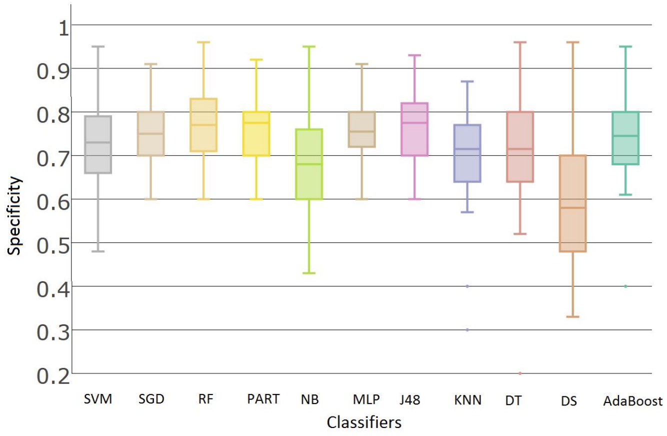

Fig. 4. Specificity for Different Classifiers Obtained For 38 Drivers.

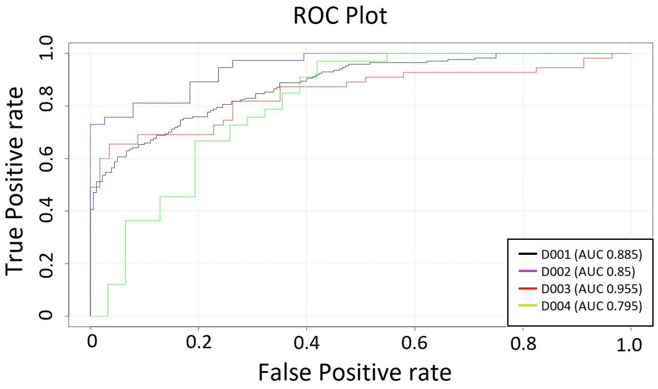

Fig. 5. ROC plot showing True Positive Rate and False Positive Rate for 4 drivers

classifier. Fig. 2 provides a representation of the variation in accuracy, for different classifiers, across all the drivers. It is clear that Random Forest is the best classifier and also the only classifier to have a median accuracy greater that 80%. Table 2 provides median and standard deviation of accuracy for different classifiers. Adaboost is the second best classifier; SVM, SGD, MLP and Adaboost perform almost equally well. Each of the obtained median accuracy is higher than 70%, thereby demonstrating the effectiveness of custom (i.e. driver specific personalized) feature set. From table II, it is clear that Random Forest is suitable for authentication problem. To validate this further, we checked sensitivity and specificity for all these classifiers. For driver authentication problem, apart from accuracy figure, sensitivity and specificity should be equally high. Fig. 3 shows sensitivity values for different classifiers used to authenticate drivers. Random Forest gives median sensitivity of 0.86. Decision Stump shows best sensitivity of 0.94 and SVM performs well in comparison to Random Forest. But we need to check specificity also for Decision Stump and SVM.

Comparing all classifiers random forest turns out to be the best. To check robustness of the model ROC [12] is analyzed for random forest. ROC plot for 4 drivers is presented in Fig 5 along with their area under ROC curve (AUC) value. ROC analysis for drivers with all features instead of personalized features were performed. The corresponding result is presented in Fig. 6. We get better performance with computational efficiency as only 40 features are needed to be computed instead of 137 features. Obtained AUC (Area under ROC curve) value for each driver is presented as a box plot in Fig

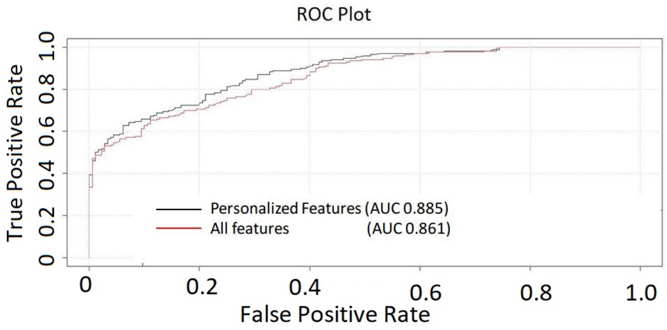

Fig. 6. ROC plot showing True Positive Rate and False Positive Rate for Driver D001 with all features (137) and personalized features (40) showing superior performance with improved AUC value

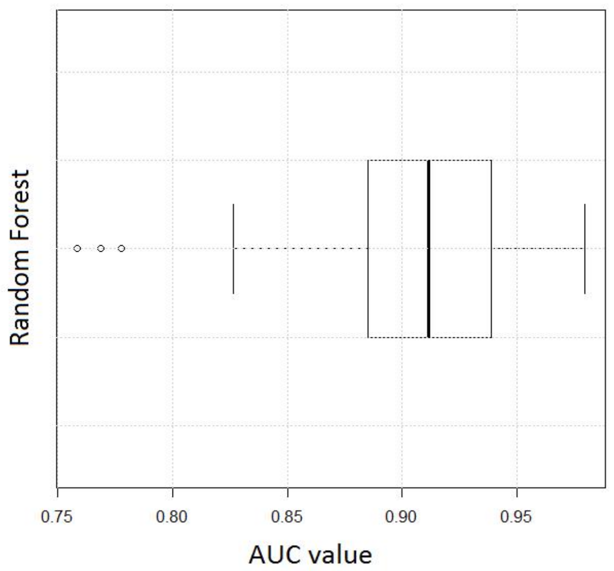

Fig. 7. Box plot of AUC obtained for all drivers using Random Forest Classifier

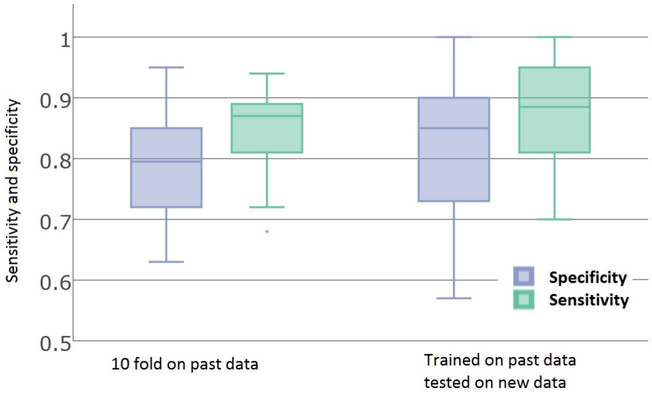

Fig. 8. Specificity and Sensitivity for 10 Fold Cross Validation and on Test data using Random Forest

7. Mean AUC obtained is 0.9 and maximum is 0.98 (for driver D014) and minimum 0.75 (for driver D061). Only for three drivers AUC is below 0.8. Thus AUC analysis confirms that proposed method is robust. Hence Random Forest is finalized as the best classifier and further analysis is done.

The validity of the proposed personalized feature set is checked by training our models on 80% of total data (1st 7 weeks) and testing on 20% (last 2 weeks) using Random Forest. Fig. 8 shows comparison of sensitivity and specificity using Random Forest applied on test data and 10 fold cross validated data. It is seen that sensitivity and specificity improves on test sets compared to 10 fold cross validation. The same is true for accuracy. Median accuracy for 10 fold is 83% whereas on test data its 86.93%. The implemented system using Random Forest with personalized feature set works well for drivers.

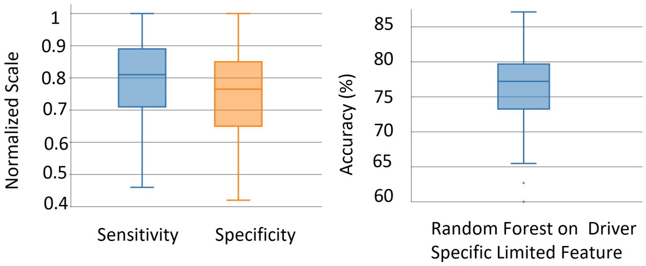

Fig. 9. Specificity, Sensitivity and Accuracy for 2 weeks data (20%) as training data and last 7 weeks (80%) data as testing data using Random Forest on Driver Specific limited feature.

In real life, training data availability may be less. Thus, it is important to see how the proposed system will work with less amount of training data. In order to investigate that, the machine learning model is trained on 20% data(1st 2 weeks) and tested on remaining 80% data. Results are found to be promising as shown in Fig. 9. Clearly, sensitivity is higher than specificity with median (of both) crossing 0.75. Even with limited training data, the median accuracy crosses 75%. Table III summarizes median and standard deviation of specificity, sensitivity and accuracy for the following cases:

(i) 10 fold cross validation on data(initial 7 weeks)
(ii) Initial 7 weeks of training (80% of the data) and last 2 weeks (20%) testing
(iii) Initial 2 weeks of training (20% of the data) and last 7 weeks (80%) testing.

Median is chosen as a proper representation as it is a robust measure of central tendency and less prone to outliers than mean [22].

Table III and Fig 9 shows that the proposed solution gives good and steady performance with driver specific personalized feature set based only on a single sensor (GPS) data. Sensitivity, specificity decreases by a small amount when initial 2 weeks data are used for training and last 7 weeks data are used for testing as compared to initial 7 weeks data for training and last 2 weeks data for testing. Here, accuracy decreases but the standard deviation of accuracy is also reduced implying more consistency. This again highlights that the personalized feature sets are relevant and the proposed method displays a kind of fingerprinting of driving style. Results are effective for 10 fold cross validation on initial 7 weeks data. Robustness is further confirmed by AUC values obtained (Fig 7).

IV. Conclusion and Future Directions

The proposed method achieves more than 80% average accuracy for authentication of a driver based on single sensor (i.e. GPS) to a good extent. This is at par with hidden Markov model (HMM) based methods on multi sensor approach [5]. Enev et al. [4] found Random Forest to be best classifier (among 4 classifiers of kNN, Naïve Bayes, SVM and Random Forest) while detecting identity of driver in a multi sensor

TABLE III
Accuracy Sensitivity and Specificity Measure and Variation for Conducted Experiment

Experimental	Sensitivity		Specificity		Accuracy (%)	
Cases	Median	StandardDeviation	Median	Standard Deviation	Median	Standard Deviation
(i)	0.87	0.12	0.79	0.12	83.01	12.01
(ii)	0.88	0.09	0.85	0.17	86.93	11.23
(iii)	0.81	0.19	0.76	0.13	77.21	8.32

approach. Random forest also emerged as best classifier for driver identification (in a group of 4-6) in [6]. Our research reaffirms their findings. This model will only for work drivers whose driving data has already been stored in the database. For addition of new drivers, training phase has to be re-executed.

This technique preserves privacy and works with minimal sensing. Deployment is very easy and scalable. Modern smartphone or any other embedded device placed in the car can execute this algorithm without manual intervention. For future, authentication can be improved further by adding another level of authentication on top of this method. Moreover, here we had less than 300 trips per driver, so only machine learning approach is used. For higher cardinality of dataset, supervised deep learning algorithms can be used for better results.

References

[1] M. Phelan, "Driver authentication system and method for monitoring and controlling vehicle usage," Apr. 9 2013, uS Patent 8,417,415.

[2] S. Ota, "Apparatus for authenticating vehicle driver," Sep. 28 2006, uS Patent App. 11/384,678.

[3] J. Wahlström, I. Skog, and P. Händel, "Smartphone-based vehicle telematics: A ten-year anniversary," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 10, pp. 2802–2825, 2017.

[4] M. Enev, A. Takakuwa, K. Koscher, and T. Kohno, "Automobile driver fingerprinting," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 1, pp. 34–50, 2016.

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

[6] A. Chowdhury, T. Chakravarty, A. Ghose, T. Banerjee, and P. Balamuralidhar, "Investigations on driver unique identification from smartphones gps data alone," *Journal of Advanced Transportation*, vol. 2018, 2018.

[7] S. Ezzini, I. Berrada, and M. Ghogho, "Who is behind the wheel? driver identification and fingerprinting," *Journal of Big Data*, vol. 5, no. 1, p. 9, 2018.

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

[9] R. C. Team *et al.*, "R: A language and environment for statistical computing," 2014.

[10] M. B. Kursa, W. R. Rudnicki *et al.*, "Feature selection with the boruta package," *J Stat Softw*, vol. 36, no. 11, pp. 1–13, 2010.

[11] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.

[12] T. Fawcett, "An introduction to roc analysis pattern recognition letters 27, 861–874," 2006.

[13] I. Barandiaran, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, 1998.

[14] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[15] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," 1998.

[16] D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, and B. W. Suter, "The multilayer perceptron as an approximation to a bayes optimal discriminant function," *IEEE Transactions on Neural Networks*, vol. 1, no. 4, pp. 296–298, 1990.

[17] J. Girones, "J48 decision tree," http://data-mining.business-intelligence.uoc.edu/home/j48-decision-tree, Accessed 13 November 2018.

[18] B. Lantz, *Machine learning with R*. Packt Publishing Ltd, 2013.

[19] R. Kohavi, "The power of decision tables," in *European conference on machine learning*. Springer, 1995, pp. 174–189.

[20] N. M. Nasrabadi, "Pattern recognition and machine learning," *Journal of electronic imaging*, vol. 16, no. 4, p. 049901, 2007.

[21] "Class adaboostm1," http://weka.sourceforge.net/doc.dev/weka/ classifiers/meta/AdaBoostM1.html, Accessed 13 November 2018.

[22] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, 2013.