Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Predicting Traffic Phases from Car Sensor Data

Shammy Uniyal[a], Emiliano Heyns[a]*, Frans Tillema[a], Chris Huijboom[a]

[a]*HAN University of Applied Sciences Automotive Research, Ruitenberglaan 29, 6826 CC Arnhem, the Netherlands*

Abstract

This research is an explorative study to look for the potential to predict traffic density from driver behavior using signals collected from the Controller Area Network (CAN) bus. The hypothesis is that driver behavior is influenced by traffic density in such a way that an approximation of the traffic density can be determined from changes in the driver behavior. Machine learning will be employed to correlate a selection of commonly available sensors on cars to the traffic density. Challenges in the processing of the data for this purpose will be outlined. The data for this study is collected from five passenger cars and nineteen trucks driving on the A28 highway in Utrecht region in the Netherlands. This study is restricted to straight roads in order to isolate the steering behavior attributable to the traffic state influences rather than following the curve in the road. The results are encouraging that the correlation between driver behavior and traffic density can be established. An overall accuracy of over 95% is achieved with a precision of 92%. The recall rate however is low most likely caused by over-fitting due to the unbalanced dataset. The results still look promising and more training data should improve the results.

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY.

Keywords: Congestion; Traffic Density: Driver Behavior; CAN Bus; Probe Vehicle Data; Supervised Machine Learning

1. Introduction

Traffic jams and road safety problems are rising with the increasing number of vehicles on the road. Traffic congestion leads to negative health impact, waste of fuel and unproductive hours (Condurat et al. (1)). In 2017, traffic congestions cost an estimated total of €80 billion in the European Union and $305 billion in the United States. So, tackling traffic congestion is imperative to reduce wastage of energy and resources. With the advance of technology, we can now detect jams effectively. It would be better however to predict them in advance, so preventative measures can be deployed. This paper presents the results of using the driver behaviour signals obtained via commonly available

* Corresponding author. Tel.: +31 6 46836010.
E-mail address: Emiliano.Heyns@han.nl

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY

vehicle sensors to detect the traffic phase that is likely to precede congestion. Driver behaviour here refers to direct driver input to the vehicle.

The driver behaviour would be speed changes, operation of throttle and brake pedals and steering angle changes. The driver behaviour is hypothesized to correlate significantly with the traffic density. Under the three-phase traffic theory by Kerner (2), traffic flow knows three recognizable phases of traffic density (free flow, synchronised flow and wide-moving jam) which typically follow each other as the traffic density rises. If a detectable behaviour correlates with these ranges in the traffic density, a detection of behaviour associated with the synchronized flow could function as an early warning for the wide-moving jam phase. This study intends to use driver behaviour changes to detect/predict traffic density. For this study the road under consideration would be a highway.

There has been significant research in the field of driver behaviour (Fugiglando et al., Ferreira et al., Teja,Wang et al., Li et al. (3–7)). Most of this research aims at improving driver behaviour and increasing safety. There have been fewer studies which directly relate to this work and during the literature research work for this study, one such research by Ito and Kaneyasu (8) was found. (8) uses driver behaviour to predict traffic phase or state. Driver behaviour is influenced by traffic conditions and surroundings (Condurat et al., Teja, Ito and Kaneyasu, Ma et al. (1, 5, 8, 9)). Li et al. (7)) show how highway driving is characterized by both longitudinal and lateral manoeuvres. The lateral behaviour, mainly high steering rate, can be a reflection of gap-seeking behaviour indicative of rising traffic densities. Ito and Kaneyasu (8) use neural networks to predict traffic congestion from driving input signals (driver behaviour) with an average accuracy of 81.65% in detecting traffic phases. In this study, we will try diverse means of machine learning to see if different approaches yield better detection rates. The traffic density to estimate traffic phase is obtained using openly available traffic information data from NDW (Dutch National Data Warehouse for Traffic Information).

Driver behaviour is of course not the only factor in the occurrence of congestion. Traffic is also affected by rush hours, type of road, and the weather. To isolate the driver behaviour component, for this study we have chosen a stretch of highway with few on/off ramps and no curves. This helps in minimizing the effect of road interference and intersections and minimizes the assumption errors arising due to instantaneous changes in traffic density due to on/off ramps.

Nomenclature

2D-IIDM	2D-Improved Intelligent Driver Model
ACC	Active Cruise Control
ANPR	Automatic Number Plate Recognition
CAN	Controller Area Network
FCD	Floating-Car Data
GPS	Global Positioning System
NDW	Dutch National Data Warehouse for Traffic Information
PCA	Principal Component Analysis
rica	Reconstruction Independent Component Analysis
sparsefilt	Sparse Filtering

2. Data Procurement and Analysis

The data for this research has been taken from data collected by the company SD-Insights in the Netherlands between the years 2015 and 2018. A selection was made for cars driving on a 12 km section of the A28 highway (see Figure 1) that was frequented by multiple SD-Insights cars at multiple times. This is a straight road section with just two major exit points in between. We chose this section because the major influence on driving behavior and traffic

Figure 1: A28 Highway selected for this research

density would be the number of cars on the road rather than perturbations from on/off ramps and curve-following. The A28 is one of the major motorways in the Utrecht region of Netherlands and it uses inductive loops, radar, Bluetooth, ANPR and FCD to measure traffic information such as flow rate and average speeds. The traffic information is openly provided by Dutch National Data Warehouse for Traffic Information (NDW).

The trips selected for this research spans January 2018 to March 2018. The selection includes a total of 24 separate vehicles; 5 passenger cars (27 trips across the selected road section) and 19 trucks (100 trips across the selected road section). All data was sampled at 10Hz. The initial goal was to mainly use data from passenger cars, but SD-Insights monitors primarily trucks. The collected data contained sensor data read from the CAN bus but did not contain driver demographics or make/model of the vehicle for reasons of privacy and security. It is known that driver characteristics such as age and gender have statistically significant influence on driving behavior (Feng et al. (10, p. 130)). While it is our expectation that adding this information to the dataset would further increase the accuracy of the flow phase detection, such information is not available as car sensor data, so it is not included.

Table 1. In-Car Signals Available in Data.

Signal	Units	Selected for traffic detection
Timestamps	date and time to second precision	Yes
Global Positioning System (GPS)	latitude/longitude in degrees	Yes
Speed	km/h	Yes
Brake	0 for no brake, 1 for brakes applied	No
Wipers	0 for no wipers, 1 for wipers on	No
Throttle	percentage depressed	Yes
Brake Pedal	percentage depressed	No
Steering Angle	position	No
Acceleration	m/s^2	Yes
Yaw Rate	degrees/second	Yes
Turn Indicator	0 for off, 1 for on	No

The selection aims to capture direct driver input without the use of data that represents the same information twice. Brake use is comparatively rare on highways (Feng et al. (*10, p.128*)); brake signals are also not available for all vehicles, so brake behavior is not used in this study. Throttle is selected as throttle use is expected to differ between free flow and jam phases, as we expect the stop-and-go characteristic typical of the jam phase to show up in the data. Also, the throttle signal gives us information on the use of Active Cruise Control (ACC), commonly in use on highways. But this cannot be said with any certainty for all vehicles as we did not have detailed information on the participating vehicles. The steering angle signal is not available for all vehicles, so yaw rate is selected as an alternate. Wipers and turn signals are not used in this study. Future phases of this study could take weather conditions into account as inferred from the ambient temperature, wipers and foglamp data available on the CAN bus. Additionally, we feed the model the vehicle type (passenger car vs truck); as we will later see, these exhibit very different behaviors.

A first analysis shows the vehicle data occasionally has gaps where for a period of time, data was either not available or not stored. This could be because of loss of signals or sensor malfunction. These gaps in the data will affect the accuracy of the outcomes, especially since some of these gaps occurred in the density regions of interest (synchronized flow/wide moving jam). Twelve of the 127 trips have multiple gaps. The gaps range from a few seconds up to one hundred seconds. This could lead to potential over-fitting as some vital information about phase changes could be missing.

The figures below show the speed, acceleration and yaw rate for a selection of trips (shown here to highlight different types of behavior) of passenger cars and trucks. Figures 2 and 3 show behavior of cars on the selected highway section; Figures 4 and 5 show the same for trucks. Just by visual inspection, it can be seen that the signal patterns and limits are different for cars and trucks. The speed range difference is due to different speed limits for heavy vehicles but the difference in yaw rate and acceleration between cars and trucks can also be observed.

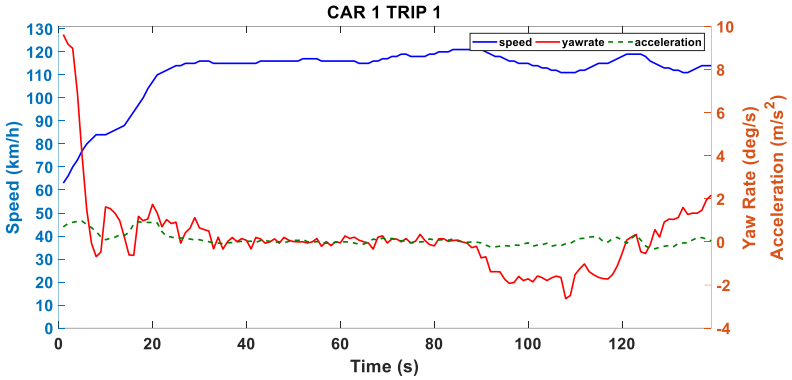

Figure 2: Speed, Yaw and Acceleration of Car 1 Trip1

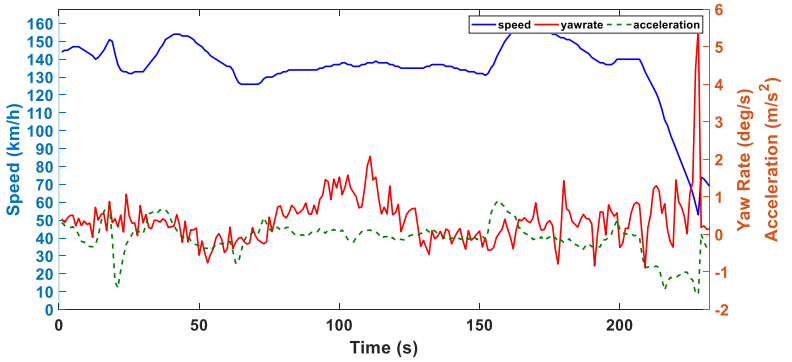

Figure 3: Speed, Yaw and Acceleration of Car2 Trip2

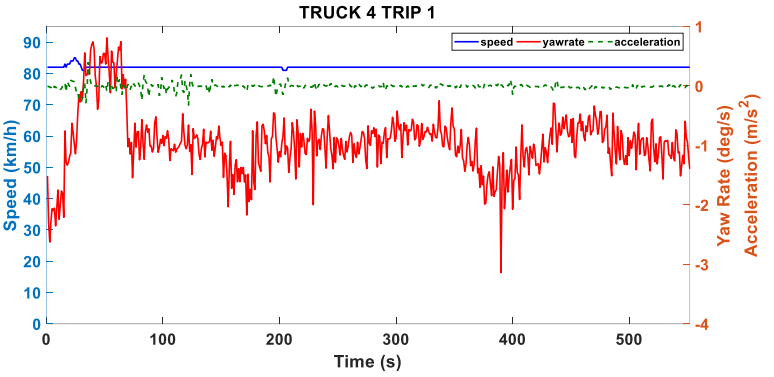

Figure 4: Speed, Yaw and Acceleration of Truck4 Trip1

First visual inspection leads us to believe a correlation between traffic state and driver behavior can be found in these sensor data. For example, Figure 5 could indicate a traffic breakdown.

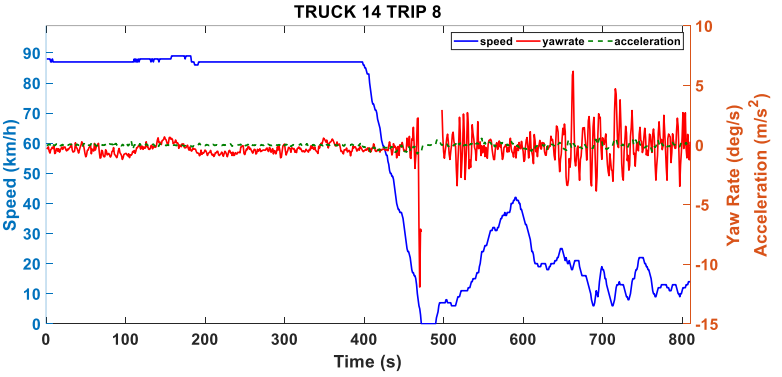

Figure 5: Speed, Yaw and Acceleration of Truck14 Trip8

2.1. Traffic Phase from NDW Data

The traffic information data for determining traffic density is obtained from the Dutch National Data Warehouse for Traffic Information (NDW) (National Data Warehouse for Traffic Information (*11*)) for the Dutch highway network. NDW uses loops, radar, Bluetooth, ANPR and FCD to record traffic data. It provides open data which includes timestamped traffic flow (q) and average traffic speed (v); from the flow and average speed we can derive the traffic density (k) as

$$k = \frac{q}{v} \tag{1}$$

On the selected road section each loop is approximately 400 m apart, for each car measurement we take the closest loop to be representative of the traffic condition at that time; depending on vehicle speed this means the loops are at most 60 seconds away from each other. Because of the spacing of the loops, not every timestamped car measurement had a corresponding loop measurement; such gaps in the loop data were filled in by extrapolating from the last-known value. If a gap of over 30 seconds occurs at the beginning or the end of the data, they are removed.

All loop data corresponding to each vehicle is combined and the fundamental diagram in Figure 6 shows the combined traffic flow diagram (per lane) for all trips under observation. Synchronized flow is visible between k1 and k2. In this diagram the density values for the lanes are averaged out as we do not know in which lane the vehicles travelled.

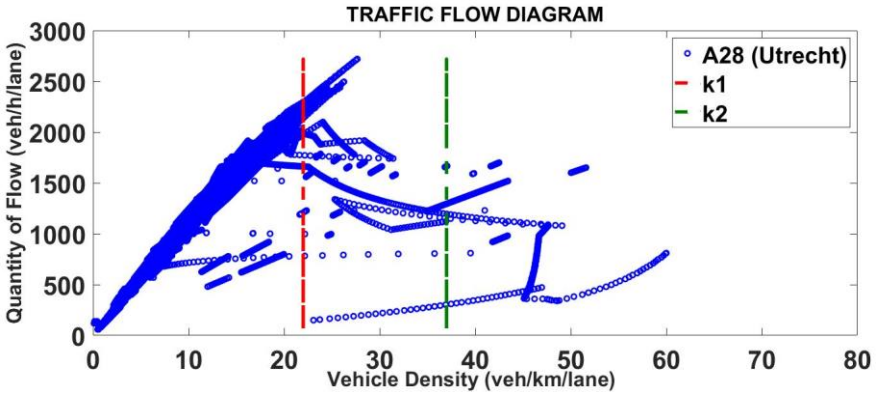

Figure 4: Traffic Flow Diagram

The fundamental diagram highlights the approximate region where different phases may lie empirically but finding this exact region by just traffic flow and traffic speed is not possible. The real traffic densities for phase changes will differ per the type of road and number of lanes on the road. In order to theorize about the traffic behavior for a particular road, more information is needed, or assumptions must be made.

For this explorative study, results from traffic simulation using Improved 2D Intelligent Driver Model (2D-IIDM) (Tian et al. (12)) are used as reference; the density range between 22 and 37 vehicles per kilometre per lane is deemed synchronized flow by Tian et al.(12, p.7) as shown in figure 6; anything below 22 is deemed free flow, and anything above 37 is deemed wide moving jam. The results from Tian et al. (12) are for homogeneous traffic; the values for k1 and k2 will differ for real heterogeneous traffic. This can be observed in a 2D-IIDM-based simulation†. Knowing this, using the values from Tian et al. (12) will introduce an unknown error in determining exact phase change point based on density as some assumptions or parameters from Tian et al. (12) may not apply on data used in this study. At later stage or further research, the actual critical densities for the road sections under study must be calculated from the NDW data, but time constraints prevented inclusion of these in the current study.

With the ranges so defined, the occurrence of traffic phases in each trip is determined to check the distribution of data points across each traffic phase. It is found that the number of data points for free flow are overwhelmingly more than those for synchronized and wide moving jam. This distribution gives a high chance of over-fitting as the model will learn more about free flow and will not be able to accurately detect other phases given new data.

3. Data Analysis and Pre-processing of Input Signals

Since we cannot assume the sensors to be fully reliable, the first step is to select the trip data for training the model. To this end, all data files are checked for consistency of data: whether they contain all required signals, and whether there are too many missing values. The throttle signal is normalized as different vehicles have different throttle pedal position ranges:

$$throttle' = \frac{throttle - throttle_{min}}{throttle_{max} - throttle_{min}} \quad (2)$$

† http://www.traffic-simulation.de/ by Treiber and Kesting(13)

Figures 7 to 10 show that the number of occurrences of acceleration, speed, throttle and yaw behavior show different patterns in different phases, which would indicate that machine learning should be able to find these correlations.

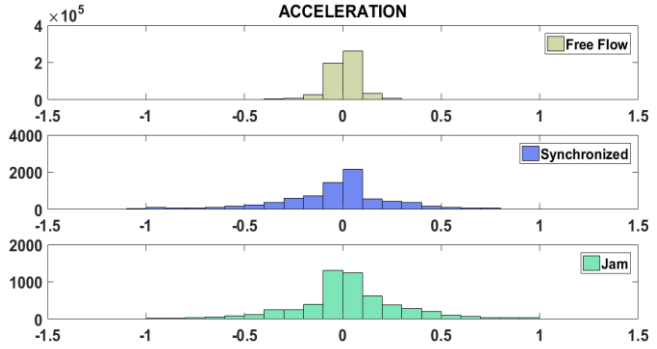

Figure 5: Number of occurrences of acceleration values for different traffic phases

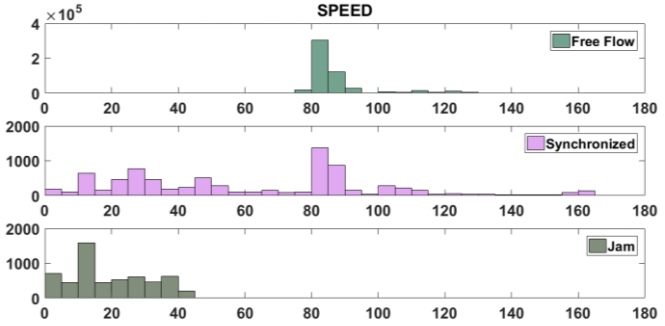

Figure 6: Number of occurrences of speed values for different traffic phases

Figure 9 shows more variance for throttle in the jam phase as there is continuous use of the throttle pedal to close any gaps as they occur. The value 0 for throttle in free flow and synchronized flow may be due to the use of ACC in most vehicles, but for jam the value 0 is expected as the vehicles would stop and go in congested traffic.

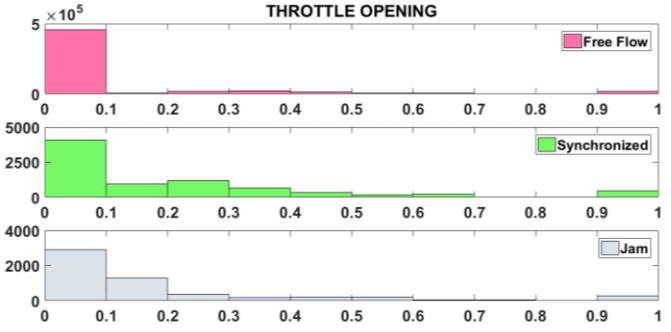

Figure 7: Number of occurrences of throttle values for different traffic phases

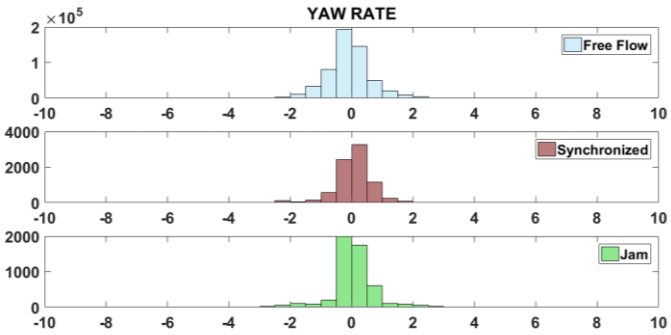

Figure 8: Number of occurrences of yaw rate values for different traffic phases

In order to optimize detection accuracy further, feature extraction is used to add more features for the classification learner. This is done using Reconstruction Independent Component Analysis (rica) and Sparse Filtering (sparsefilt), which can be used for feature extraction. The number of features were selected by increasing features one by one until there was no significant increase in accuracy as compared to the increase in training time with higher dimensionality. Using this procedure, a total of six new features are extracted to avoid increasing dimensionality too far which would significantly increase training time. ‘rica’ creates a linear transformation of input features and 'sparsefilt’ creates a non-linear transformation of input features to output features. Three sets of training data are prepared. One with initially selected input features and two other prepared using ‘rica’ and ‘sparsefilt’.

4. Training of Predictive Model

In this study, we use an ensemble learning approach using Bagged trees and RUSBoosted trees. RUSBoosted trees are recommended for imbalanced datasets (Seiffert et al. (14)). Accuracy is improved by testing different configurations for the Bagged and RUSBoosted algorithms. The configuration includes parameters such as learning rate, maximum number of learners, number of splits. Input features are removed or added, and the results are checked for improvement. The parameters of the ensemble algorithms and their effects on training are described below:

- The learning rate was varied between 0.1 to 1. A learning rate of 0.1 would take longer to train but would usually achieve higher accuracy.
- The maximum number of splits control the depth of tree learners. The number of splits was gradually increased as too many branches can lead to over-fitting.
- The number of learners can be increased to achieve higher accuracy at the cost of longer training times.

The parameters are tuned gradually to avoid over-fitting and keep the training time as low as possible. Multiple rounds of training are conducted with different configurations and algorithms. Besides the ensemble algorithms, other algorithms are also tried. All the results from these algorithms are analysed and compared. The algorithms which produced the best results are trained again with multiple configurations. Each configuration is also trained after performing Principal Component Analysis (PCA) on input features to see if that improves results. This is repeated until the best results are achieved from a specific configuration, with minimum training time.

The same process is repeated with the other two datasets created using feature extraction. The extracted features are also normalized during feature extraction, so all algorithms which perform better after normalization are also used for training to check if higher accuracy is achieved. Different algorithms have different configuration options which are also changed, and multiple training simulations are run. They are not described here as they produced low accuracy results and ensemble algorithms still performed better. So, ensemble algorithms are trained multiple times with different configurations to achieve best possible results.

5. Testing and Validation

The testing and validation were performed in MATLAB which provides three different techniques for validation: Cross Validation, Hold-Out Validation and No Validation. Since the data set in this study is not very large, cross-validation is used for best results. Moreover, since the dataset for this study is severely unbalanced, cross-validation helps in reducing over-fitting.

6. Results

Three datasets are used for training. The first set is the one obtained after combining all selected input features (speed, yaw rate, acceleration and throttle). Speed is the major predictor as it explains most of the differences between all inputs. PCA was used and 99.7% variance could be explained with first principal component, but the results did not improve by using PCA. The other two sets are prepared by using functions rica and sparsefilt.

Many training simulations were run using all three datasets. The ensemble algorithms (Bagged and RUSBoosted Tress) are used. Other algorithms were also tried but their results were not significant (very low accuracy), therefore they are not described here. As expected the imbalance in the data did affect the results. Figures 11 and 12 show scatter plots of two predictors with different colors indicating different traffic phases of first and second dataset respectively. It is clearly visible that the data points for free flow (blue colored) are more in number.

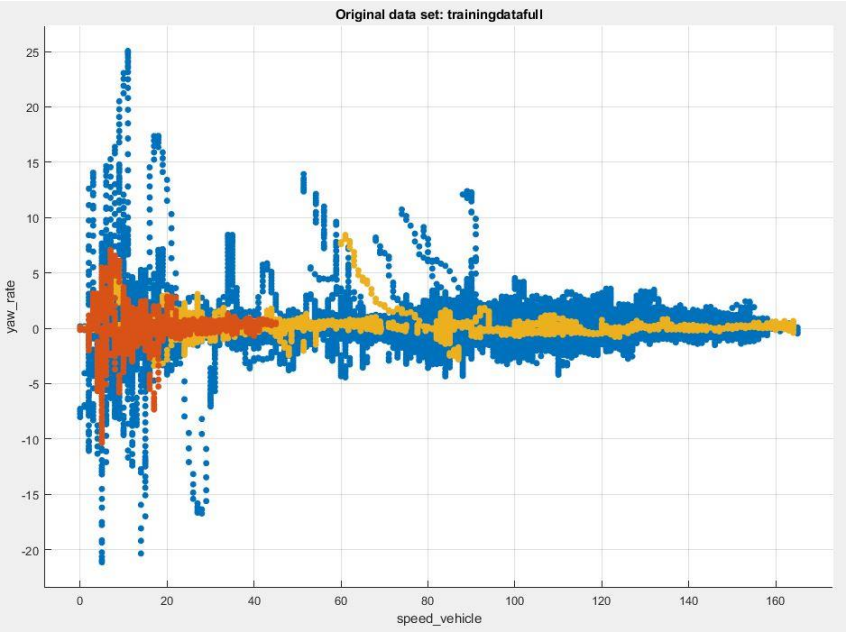

Figure 9: Scatter plot of predictors speed and yaw rate for Dataset 1. Blue-Free Flow, Yellow-Synchronized Flow, Orange-Wide Moving Jam

The ensemble algorithms were tuned using different parameter values to achieve the best possible results given the dataset available. The most optimum results of datasets are formulated in Table 2. The results of third dataset

(sparsefilt) have been omitted here as they displayed very poor results. Precision or positive predictive rate indicates the accuracy with new data or validation data. Recall rate or true positive rate indicates accuracy of training data.

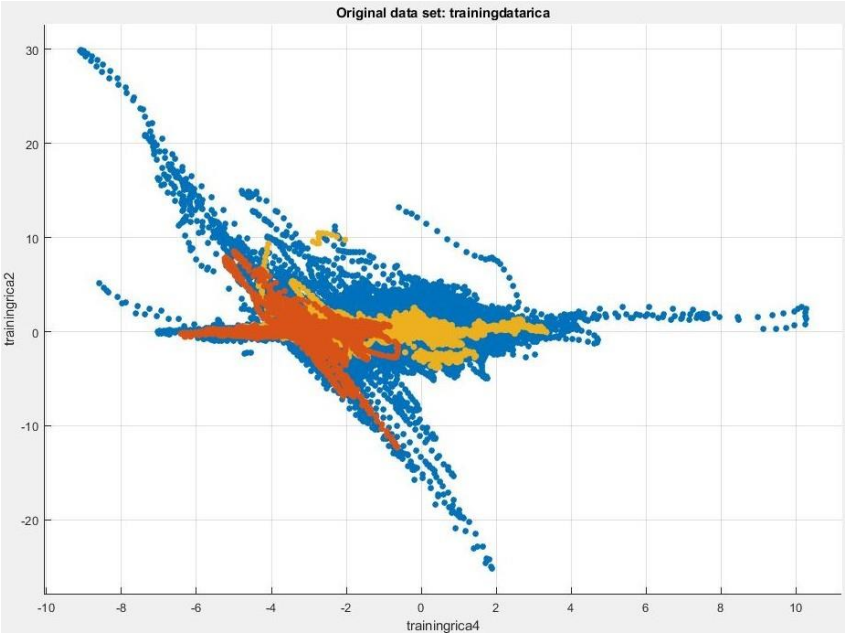

Figure 10: Scatter plot of two predictors for Dataset 2. Blue-Free Flow, Yellow: Synchronized Flow, Orange-Wide Moving Jam

Table 2: Accuracy achieved using different datasets with precision and recall rate of synchronized flow phase.

Dataset	Algorithm Used	Configuration	Accuracy and Training Time
First Dataset	Bagged Trees	Number of Learners = 200 Maximum Number of Splits = 50000	Overall Accuracy = 99.3 % Training Time = 8053.6 seconds Precision of synchronized flow = 92 % Recall Rate of synchronized flow = 65 %
	RUSBoosted Trees	Number of Learners = 250 Maximum Number of Splits = 1500 Learning Rate = 1	Overall Accuracy = 96.7 % Training Time = 982.15 seconds Precision = 33 % Recall Rate = 82 %
Second Dataset (rica)	Bagged Trees	Number of Learners = 200 Maximum Number of Splits = 100000	Overall Accuracy = 99.2 % Training Time = 15304 seconds Precision of synchronized flow = 62 % Recall Rate of synchronized flow = 91 %
	RUSBoosted Trees	Number of Learners = 400 Maximum Number of Splits = 2000 Learning Rate = 1	Overall Accuracy = 98.7 % Training Time = 2491.4 seconds Precision of synchronized flow = 64 % Recall Rate of synchronized flow = 75 %

For the first dataset using bagged trees, it is observed that the recall rate is low for synchronized flow, but the precision is high. The jam phase shows high recall and precision, this is due to it being easily distinguishable by the major predictor ‘speed’. The model is unable to correctly learn the synchronized phase due to too few data points for this phase against the large number of free flow points, and over-fitting occurs. As a result, some of the data points from synchronized class are also classified as free-flow. Also, as we can see in the scatter plot (Figure 11), free flow and synchronized flow coincide at almost all points so other predictors need to be used for differentiating, but since there are so much more data points for free-flow the model by default assigns free-flow to most data points when it is confused. The large number of data points for free flow is also the reason overall accuracy is not the real performance indicator because even if the other two classes (synchronized and jam) have low number of true positives, the overall accuracy would still be high if free flow is correctly predicted. The precision of the model is pretty good; out of all data points recognized as synchronized, 92% are correctly classified. A traffic manager would likely prefer high precision, so that the chances of unnecessarily diverting traffic due to false detection of synchronized flow are low.

Using RUSBoosted trees algorithm with first dataset did improve the recall rate of synchronized flow phase but it also decreased the precision; that is, a greater number of synchronized flow points were correctly predicted during training. This is the result of random under sampling of the free flow class, balancing the dataset. But this also results in less training of free flow data points and in turn the model predicts more free flow points as synchronized. Therefore, the precision of the model decreases, and the model performs poorly with new data due to over-fitting.

For the second dataset (rica), using bagged trees did not improve the results (Table 2) over the original dataset and the training time also increased due to higher dimensionality with six features. Using RUSBoosted trees however, did improve the precision but decreased the recall rate as compared to dataset 1. This indicates that the model performed a little better but still lacks some information needed which could be corrected with more sample data.

Figure 13 shows the scatter plot of two predictors of third dataset (sparsefilt). It can be observed that non-linear transformation using ‘sparsefilt’ spread the free-flow values across the plot. This certainly would affect the result and the models trained using this dataset exhibited very poor results. For this reason, the results have not been discussed here.

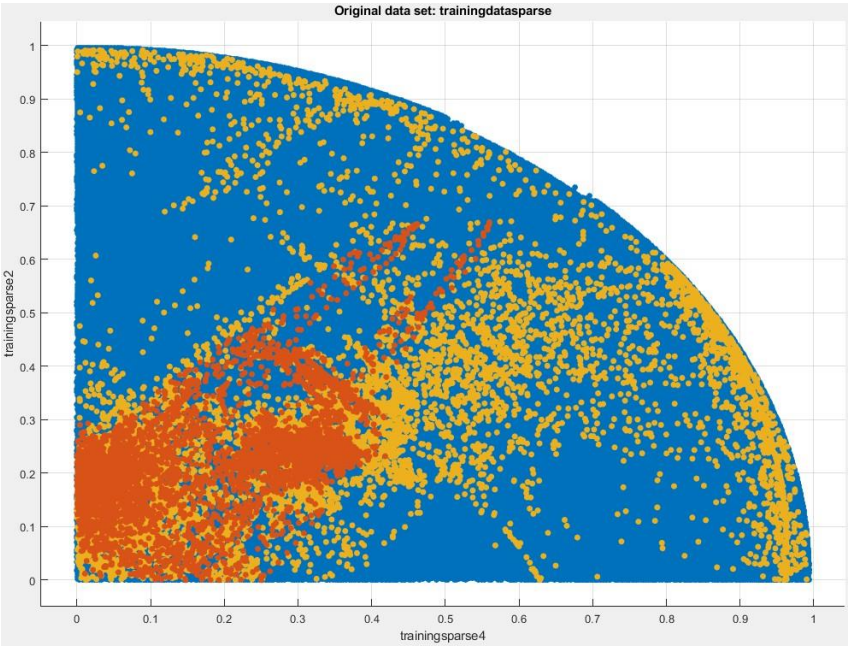

Figure 11: Scatter plot of two predictors for Dataset 3. Blue-Free Flow, Yellow-Synchronized Flow, Orange-Wide Moving Jam

Figures 14 to 16 show the confusion matrices for results of first dataset using bagged for reference. Figure 14 shows the main confusion matrix indicating the number of correctly predicted observations. Figures 15 and 16 show the True Positive Rate (Recall Rate) matrix and Positive Predictive Rate (Precision) matrix respectively.

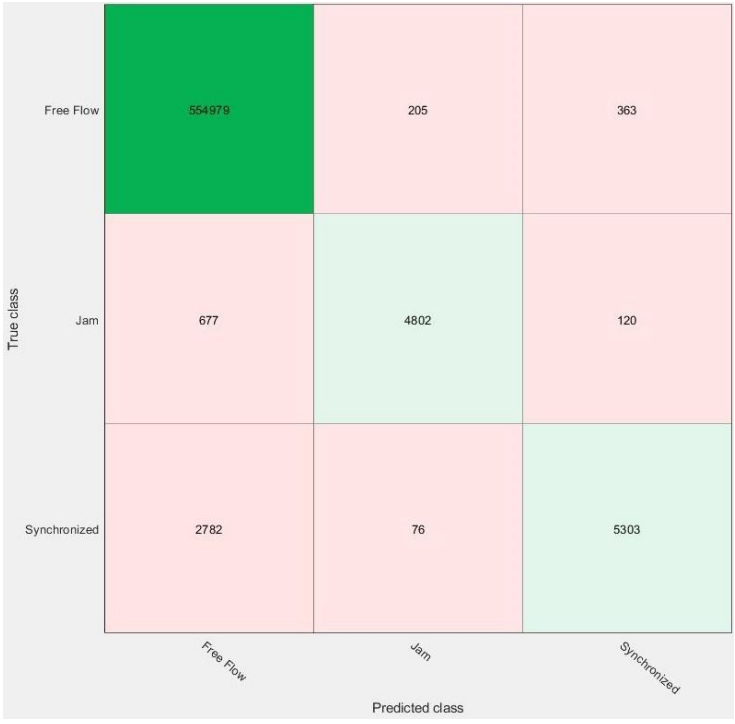

Figure 14: Confusion Matrix using Bagged Trees for Dataset 1

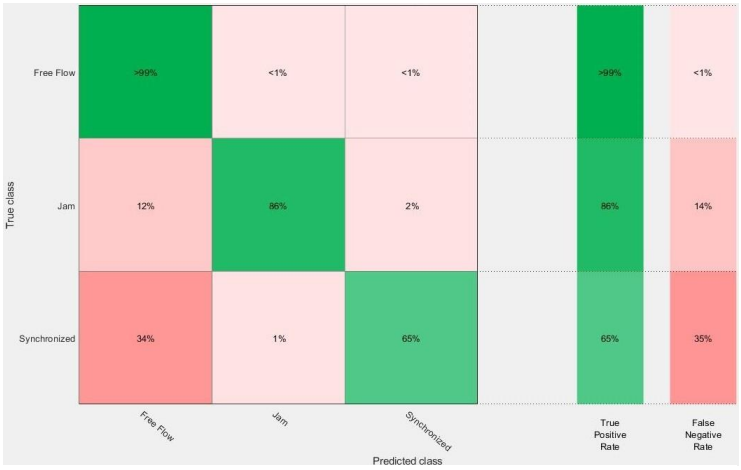

Figure 15: True Positive Rates using Bagged Trees for Dataset 1

Figure 16: Positive Predictive Values using Bagged Trees for Dataset 1

6.1. Training Results Discussion

Even though the 'speed' input largely dominates the other inputs, the other input features still relevantly influence the classification, as removing them decreases accuracy. The imbalance in the data is very evident in the results, but the model is still able to classify with reasonable accuracy (high precision). More data, specifically from congestion phases, is needed to further improve classification results. Extracting more features does not significantly help as there is no significant improvement in classification performance. This could mean that the maximum possible performance is achieved given the original dataset. High recall and precision could not be achieved simultaneously. This is caused by the insufficient quantity of data. The high precision with bagged trees shows that the model performs reasonably well with new data. Although high recall (sensitivity) is achieved using RUSBoosted trees, it fails to perform well with new data, as the precision decreases. So, there is trade-off here between recall and precision. For this problem, high precision seems more desirable as the objective is to detect synchronized phase as a predictor (precursor) to the jam phase. With high recall many free flow data points are also recognized as synchronized, which would lead to wrong prediction of the jam phase, that is, if a model such as this is implemented, and traffic is redirected based on this model then there is a high chance that the model detects free flow as synchronized and traffic is unnecessarily diverted.

7. Limitations

The data collected came mostly from trucks (19) compared to passenger vehicles (5). 87.5% of the collected data is of trucks. This means the model as it is currently trained will be more effective in detecting traffic states through

truck behavior data. The model could be improved by retraining it with more data from passenger cars, as this would better reflect the typical traffic mixture on highways. The traffic states for training of the model have been approximated by sing results of a a 2D-IIDM-based traffic simulation. Using the actual road configuration would further improve accuracy and usefulness of the model.

8. Conclusions

The training results show that traffic phases can be classified using driving behaviour. Driver behaviour changes as the traffic phases change and these changes can be correlated to these traffic phases using machine learning. The results of this explorative study are encouraging that the correlation can be established between driver behaviour and traffic phases, but a larger dataset and more relevant data (synchronized and jam phase) is needed. The data used in this study is biased, the effects of which is evident in results. Free flow detection has high accuracy because much more data is available to describe it. Although the recall of synchronized flow of the model is low, it shows high precision, which is good for our purposes as the goal is to detect precisely this synchronized flow, and the trained model can detect it with high accuracy. That is, what the model classifies as synchronized flow is mostly correct (92% in results). High recall is also achieved using a RUSBoosted trees algorithm, but in this case the model detects many free-flow points as synchronized, which is not desired. This study thus illustrates a way to correlate driver behavior with traffic phases and presents a technique to use machine learning and detect traffic phases through driver behavior. It is concluded that with more relevant data, this technique could be very useful and more reliable.

9. Future Work

The learning model can be improved using more data specifically with more observations from non-free-flow phases. Another suggestion would be to include steering angle or steering rate input, which would more directly take direct driver input into account and might produce better results. Also, using headway as an input variable could potentially improve results, as short headway combined with variable throttle is highly likely to occur in the synchronized flow phase. Using deep learning with a larger set of data is also recommended. A planned next phase will aim to bolster these early findings with a larger data set and implement the trained model for real-time predictions on running cars. The goal is to have a more balanced dataset between trucks and passenger cars to gain more insight in the behavior differences and how they correlate to the traffic phases

References

[1] Condurat, M., A.M. Nicută, and R. Andrei, Environmental Impact of Road Transport Traffic. A Case Study for County of Iaşi Road Network. *Procedia Engineering*, Vol.181, 2017, pp.123–130.

[2] Kerner, B.S., Definitions of The Three Traffic Phases. *An Introduction to Modern Traffic Flow Theory and Control: The Long Road to Three-Phase Traffic Theory*, Springer-Verlag, Berlin-Heidelberg,2009.

[3] Fugiglando, U., E. Massaro, P. Santi, S. Milardo, K. Abida, R. Stahlmann, F. Netter, and C. Ratti, Driving Behavior Analysis through CAN Bus Data in an Uncontrolled Environment. *arXiv:1710.04133 [physics]*, 2017.

[4] Ferreira, J., E. Carvalho, B.V. Ferreira, C. de Souza, Y. Suhara, A. Pentland, and G. Pessin, Driver Behavior Profiling: An Investigation with Different Smartphone Sensors and Machine Learning. *PLOS ONE*, Vol.12, No. 4, 2017, p. e0174959.

[5] Teja, M. S. S., Driver Behavior Detection System with Inter-Vehicle Communication. *International Journal of Engineering Research*, Vol.3, No. 10, 2014, p.5.

[6] Wang, W., J. Xi, and H. Chen, Modelling and Recognizing Driver Behavior Based on Driving Data: A Survey,2014.

[7] Li, G., S. E. Li, B. Cheng, and P. Green, Estimation of Driving Style in Naturalistic Highway Traffic Using Maneuver Transition Probabilities. *Transportation Research Part C: Emerging Technologies*, Vol.74, 2017, pp.113–125.

[8] Ito, T. and R. Kaneyasu, Predicting Traffic Congestion Using Driver Behavior, 2017, Vol.112, pp.1288–1297.

[9] Ma, C., X. Dai, J. Zhu, N. Liu, H. Sun, and M. Liu, DrivingSense: Dangerous Driving Behavior Identification Based on Smartphone Autocalibration. *Mobile Information Systems*, Vol.2017, 2017, p.15.

[10] Feng, F., S. Bao, J. R. Sayer, C. Flannagan, M. Manser, and R. Wunderlich, Can Vehicle Longitudinal Jerk Be Used to Identify Aggressive Drivers? An Examination Using Naturalistic Driving Data. *Accident Analysis & Prevention*, Vol.104, 2017, pp.125–136.

[11] National Data Warehouse for Traffic Information, Historical Data. http://www.ndw.nu/pagina/en/78/database/81/historical_data/, 2018.

[12] Tian, J., R. Jiang, G. Li, M. Treiber, C. Zhu, and B. Jia, Improved 2D Intelligent Driver Model Simulating Synchronized Flow and Evolution Concavity in Traffic Flow. *arXiv:1603.00264[nlin, physics:physics]*, 2016.

[13] Treiber, M. and A. Kesting, *Traffic Flow Dynamics: Data, Models and Simulation*. Springer-Verlag, Berlin Heidelberg, 2013.

[14] Seiffert, C., T.M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, RUSBoost: Improving Classification Performance When Training Data Is Skewed. IEEE, 2008, pp.1–4.