Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2022) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

Transport Research Arena (TRA) Conference

Investigation of the time spent in dangerous driving conditions: Findings from the i-DREAMS project

Marianthi Kallidoni[a]*, Eva Michelaraki[a], Christos Katrakazas[a], Tom Brijs[b], George Yannis[a]

[a]National Technical University of Athens, Department of Transportation Planning and Engineering, 5 Heroon Polytechniou str., GR-15773, Athens, Greece
[b]UHasselt, School for Transportation Sciences, Transportation Research Institute (IMOB), Agoralaan, 3590 - Diepenbeek, Belgium

Abstract

This paper investigates how much time is spent in three levels of a Safety Tolerance Zone (STZ) for driving speed. Towards that aim, a naturalistic driving experiment was conducted and data from a representative sample of 20 Belgian car drivers were analyzed. Two classification models (i.e. Conditional Inference Trees and Support Vector Machines) and two regression models (i.e. Support Vector Machine Regression) were utilized in order to successfully predict initially the STZ levels and then the time that drivers spent in each one. The results indicated that both classification models predict STZ levels with 92% accuracy, 96% specificity and 92% recall. Moreover, regression models could explain at least 81% of the variance, but further analysis is needed to minimize errors in duration prediction.

Peer-review under responsibility of the scientific committee of the Transport Research Arena (TRA) Conference

Keywords: i-DREAMS; Safety Tolerance Zone; Real-time safety assessment; Classification; Regression.

1. Introduction

The development of the 'Safety Tolerance Zone' (STZ) is the main aim of the European H2020 project i-DREAMS†. This zone, although abstract in nature, refers to the self-regulated control of transportation vehicles by

* Corresponding author. Tel.: +30-210-772-1380;
E-mail address: mkallidoni@mail.ntua.gr
† Further general project information can be found on the website: https://idreamsproject.eu

Peer-review under responsibility of the scientific committee of the Transport Research Arena (TRA) Conference

human operators in the context of crash avoidance, and exploits task complexity indicators along with driver background factors for a continuous real-time assessment of safe driving operation. The STZ consists of three phases: Normal Driving phase, Danger phase and Avoidable Accident phase. Normal Driving refers to the phase where conditions are safe and the crash risk is low, whereas the Danger phase is characterized by changes to normal operations and an increased probability of crash. Finally, the avoidable accident phase indicates that a collision scenario is developing, but there is still time for the operator to intervene and avoid the crash.

The objective of the current study is to identify how much time is spent in the three levels of the STZ for driving speed. To that aim, the most reliable indicators of task complexity, such as time headway and distance travelled or weather conditions are going to be assessed. To achieve this object, a naturalistic driving experiment was conducted and data from 20 Belgian car drivers was utilized. In order to predict the time spent in each STZ level, two classification models, i.e. Conditional Inference Tree (CIT) and Support Vector Machines (SVMs), were developed. The obtained results were evaluated and later implement to Support Vector Regression (SVR) models in order to predict the time drivers spent in the three aforementioned levels.

The paper is structured as follows. Initially, an introduction to the problem is provided. Subsequently, a brief description of the utilized data and the methodological approach is given. Then, the results of the statistical analysis performed are presented. Lastly, a discussion on significant findings and conclusions on the modelling of the STZ and the time spent in each level are highlighted in order to assist researchers and policy-makers.

2. Background

Predicting driving behavior by employing mathematical driver models, obtained directly from the observed driving-behavior data, has gained much attention in literature. To begin with, Yokoyama and Toyoda (2015) used Support Vector Machines (SVMs) with Gaussian kernel and an analysis method of driving behaviors based on large-scale and long-term vehicle recorder data to support fleet driver management by classifying drivers by their skill, safety, physical or mental fatigue and aggressiveness. The entropy-like model and Kullback Leibler divergence model, aiming to emphasize the behavioral departure from average drivers, was proposed for the classification. Results indicated that these methods can successfully find some informative driving operation behaviors that might cause accidents and examined a large scale log of vehicle data recorder. However, the frequencies at rare bins were small with short term operation. In the proposed method, operator's geo-location and weather were not taken into consideration, while a daily review of vehicle recorder data might not have the ability to distinguish an abnormal and unsafe behavior.

SVM and k-means algorithms have also been applied to recognize normal, aggressive or risk driving style based on the trajectory risk levels (Xue et al., 2019). Specifically, Discrete Fourier Transform (DFT), Discrete Wavelet Transform (DWT) and statistical methods were adopted to extract the effective features from trajectory data to enable the driving style recognition. The results indicated that the proposed SVM method was a more appropriate method, which can be effectively used to label the driving style, by comparison with RF, kNN and Multi-Layer Perceptron (MLP) algorithms, displaying an accuracy of 91.7%, a precision of 92.8% and a recall of 81.8%. The model with machine learning algorithm helped to evaluate the collision risk on the road network with high accuracy and also provided real-time decision support to drivers, but road conditions and traffic flow level which influence driving style were not taken into consideration.

Furthermore, Sardá-Espinosa et al. (2017) used a specific kind of decision tree algorithm, called conditional inference tree, which was utilized to extract relevant knowledge from data that pertained to electrical motors. The model was chosen as the most appropriate due to its flexibility, strong statistical foundation and great capabilities to generalize and cope with problems in the data. By looking at the distributions at the leaves of the trees, it was possible to assess how coherent the models were with respect to reality. Results indicated that there were a few outlier reports at certain nodes, and it was possible to evaluate their data individually and find previously unknown inconsistencies.

Findings from another interesting study (Das et al., 2009) revealed that conditional inference forests were the most appropriate method to identify risk factors affecting crash severity on arterial corridors. The methodology applied was quite insightful in identifying the variables of interest in the database (e.g., alcohol/ drug use and higher posted speed limits contribute to severe crashes). Results indicated that the failure to use safety equipment by all passengers and presence of driver/passenger in the vulnerable age group (more than 55 years or less than 3 years) increased the severity of injuries given a crash had occurred.

3. Data collection

In order to identify the time spent in dangerous and avoidable accident phases, a naturalistic driving experiment and data from 20 Belgian car drivers was utilized during a 3-month timeframe (from 21/07/2021 to 30/10/2021). Trip data were collected from a specific subset of the population of Belgium and additional information, demographic or personal characteristics of the examined sample (e.g. gender, age, educational level) were not included in this analysis. As a consequence, this study retains a scope of macroscopic examination of driver behavior, considering the trips produced by the drivers collectively.

In-vehicle technologies include dedicated information system tools to understand driving conditions, environment, and behavior (Michelaraki et al., 2021). They are able to provide real-time interventions to car drivers in order to improve their driving behavior and promote road safety. Visual, auditory and haptic warnings or combinations of both were found to enhance driving safety (Katrakazas et al., 2020a). Furthermore, multisensory wearable modules were found to have a robust and statistically significant effect of real-time feedback on both drowsiness and driving performance ratings. Many reviews also proved that there was a strong motivation for drivers to improve their driving style, differentiate their travel behavior from aggressive to normal and reduce their degree of exposure by receiving post-trip interventions and monitoring their driving performance (Katrakazas et al., 2020b).

In this perspective, data from the Mobileye system (Mobileye, 2022), a CardioDashcam and the CardioGateway (CardioID Technologies, 2022) which records driving behavior along with GNSS signals were used. In particular, private vehicles were equipped with Mobileye and CardioDashcam in order to monitor the road and the driving process and record events for post-trip analysis. In addition, the CardioGateway was a device which was used in order to receive the status of the STZ, and it can also provide visual and sound alerts in real-time, allowing as well the identification of the driver, in a scenario of multiple drivers per vehicle. Finally, the i-DREAMS application was also available on personal smartphones, not only to monitor the smartphone usage, as an indicator of distraction, but also for post-trip feedback, to engage drivers on their performance improvement, through a gamification strategy, that includes but is not limited to rating and scores, completing the monitoring dimensions targeted by i-DREAMS platform. Figure 1 provides a depiction of the i-DREAMS state-of-the-art technology installed in private vehicles.

Fig. 1. i-DREAMS suite of technologies installed in private vehicles

Based on the above, explanatory variables of risk and the most reliable indicators of task complexity and coping capacity, such as time headway, distance travelled, forward collision warning, pedestrian collision warning, harsh acceleration or harsh braking, lighting conditions and weather were assessed. Particular emphasis was given to the duration time that each driver spends in each STZ level; thus, a new variable, namely STZ_duration, which takes into account the different levels of STZ was created. Thus, the dependent variable was the STZ_duration, divided into three levels (i.e. Normal Driving phase: 0, Dangerous phase: 1, Avoidable Accident phase: 2). A new subset with max STZ duration time and the average of the other variables was created. Some descriptive statistics, such as mean, standard deviation, maximum value, and minimum value, are provided in Table 1.

Table 1. Descriptive statistics of the analysis variables

Variables	Description	Mean	Standard Deviation	Min	Max	Sample Size
ME_AWS_hw_measurement_mean	Headway measurement (sec)	69386.55	35611.92	0.31	99999.00	1820
ME_AWS_fcw_mean	Forward collision warning	0.00	0.00	0.00	0.05	1820
ME_Car_speed_mean	Average speed (km/h)	90.86	30.00	1.94	137.07	1820
GPS_distances_sum	Distance travelled (km)	618.57	163.94	10.87	1021.80	1820
DEM_evt_ha_lvl_M_mean	Medium harsh acceleration events	0.07	0.22	0.00	1.00	1820
DEM_evt_hb_lvl_M_mean	Medium harsh braking events	0.01	0.09	0.00	1.00	1820
ME_AWS_time_indicator_median	Lighting conditions (day/night)	1.45	0.82	1.00	3.00	1820
ME_Car_wipers_median	Weather conditions (wipers on/off)	0.03	0.16	0.00	1.00	1820
ME_Car_high_beam_median	Headlight beam with a long-range focus	0.00	0.03	0.00	1.00	1820
duration	Time spent in each STZ level	352.70	1476.50	30.00	37830.00	1820
STZ	Safety Tolerance Zone for driving speed	0.22	0.46	0.00	2.00	1820

4. Methodology

After the data collection, two classification models (i.e. Conditional Inference Tree and Support Vector Machine) were developed in order to identify the three phases of Safety Tolerance Zone, i.e. Normal Driving, Danger and Avoidable Accident. The predicted STZ values are then implemented in Support Vector Regression models in order to successfully predict the time that the driver is going to spent in each level. The structure methodology along with the proposed characteristics to estimate the time spent in each of the STZ levels is shown in Figure 2.

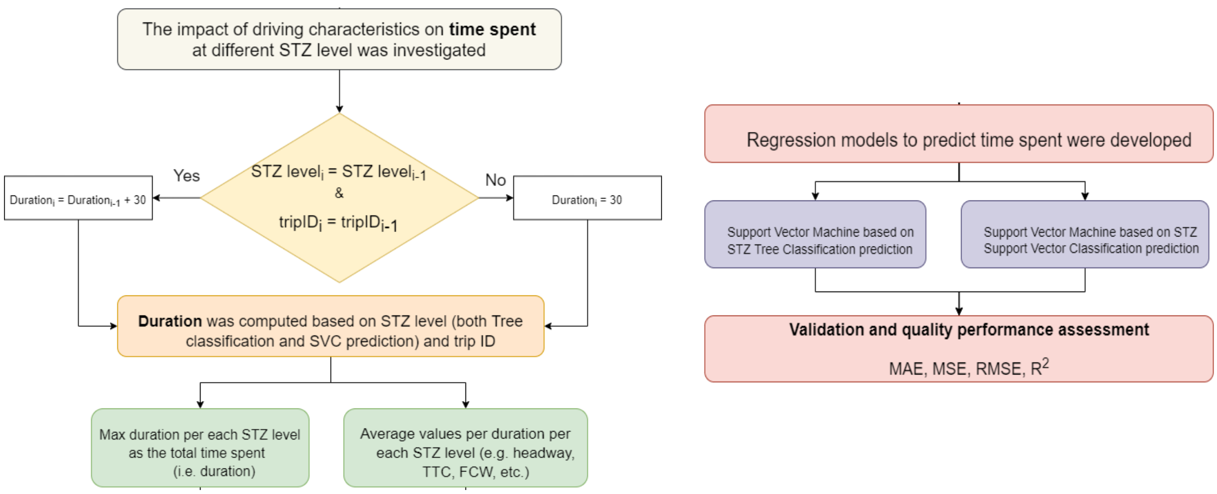

Fig. 2. Proposed methodology for the definition of the Safety Tolerance Zone for speed

3.1. Conditional Inference Tree (CIT) Classification

Decision Trees are supervised machine learning algorithms, which represent causes and effect relationships in a simplified flowchart structure, are used for both categorical (Classification Tree) and continuous (Regression Tree) data predictions. For the current classification analysis, CIT, an optimized method of the traditional CART-based trees, are employed. CITs recursively perform univariate splits of the dependent variable based on a set of covariates, similar to the CART-based trees. The main difference is that CITs introduce significant test procedures for the variable selection, known as permutations tests, in order to overcome the selection bias towards covariates with many splits or missing values (Hothorn et al., 2006).

3.2. Support Vector Classification (SVC)

Support Vector Machines (SVMs) are supervised machine learning algorithms, mostly used for classification problems. The objective of SVMs is to estimate the optimal hyperplane, i.e. the decision boundary that distinctly classifies the data points, by maximizing the margin between the support vectors of each class, i.e. the closest data points to the hyperplane. In case of non-linear mapping, a kernel trick is implemented to transform data into linear spaces (Yu & Kim, 2012). The most usual kernel parameter is the Gaussian kernel or Radial Basis Function (RBF). The parameter tuning mainly concerns the gamma (γ), which controls the width of the kernel and the cost (C), which controls the trade-off between misclassification of training examples and model simplicity (Hsu et al., 2003).

3.3. Support Vector Regression (SVR)

Except for classification problems, SVMs are also utilized to perform machine learning regressions, with the introduction of a ε-insensitive loss function, namely a penalty-free error tube around the hyperplane. The epsilon value can affect the number of support vectors (SVs) used to construct the regression function. A larger epsilon will result in fewer selected SVs, less complex regression estimates and less training time. Similar with the SVC, gamma and cost parameters control the kernel width, the model complexity and the prediction accuracy (Vapnik, 1999).

3.4. Validation and quality performance assessment

Regarding the assessment of classification models, the following metrics were utilized, based on the confusion matrix, which provides True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) metrics.

Accuracy, which is the proportion of correctly classified observations, is defined as:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

Precision, which is the proportion of true positives among all predicted positives, is defined as:

$$Precision = TP/(TP + FP) \quad (2)$$

Recall, which is the accuracy on positive examples, is defined as:

$$Recall = TP/(TP + FN) \quad (3)$$

Specificity, which is the accuracy on negative examples, is defined as:

$$Specificity = TN/(TN + FP) \quad (4)$$

F-means, which combines precision and recall into a single measure, is defined as:

$$F - means = 2 * (Precision * Recall)/(Precision + Recall) \quad (5)$$

G-means, which measures the balance between accuracies for both classes, is defined as:

$$G - means = \sqrt{Recall * Specificity} \quad (6)$$

For the assessment of regression models, the performance validity was examined using the following metrics:

The Mean Absolute Error (MAE), which gives the mean of the absolute forecasting error, is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \quad (7)$$

The Mean Squared Error (MSE), which gives the mean of the squares of the forecasting error, is defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (8)$$

The Root Mean Squared Error (RMSE), which is the square root of the average squared error, is defined as:

$$RMSE = \sqrt[2]{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (9)$$

Coefficient of determination (R^2), which is the proportion of the variation in the dependent variable that is predictable from the independent variable(s) is defined as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (10)$$

where: SS_{res} is the residual sum of squares, and SS_{tot} is the total sum of squares (i.e. proportional to the variance of the data).

It should be noted that the reason for selection of RMSE over MSE is that RMSE is a metric on a same scale as the dependent variable, instead of squared. The above metrics are easily applicable for both continuous and count data. However, metrics measuring deviation based on percentages can produce infinities given count data with zeroes in the dataset (also known as the singularity problem).

5. Results

4.1. STZ prediction

Table 2 and Table 3 provide the assessment of the two classification models, i.e. Conditional Inference Tree Classification and SVC. All analyses were conducted in R-studio (R Core Team, 2019). It should be noted that the data were split into 60% train and 40% test in order to evaluate the models. Focusing on the results of all classes combined, both classifiers achieve 92% accuracy, 96% specificity and 92% recall. Class 0 includes the majority of available data and thus was predicted with 92% precision and 100% recall, while Class 1 presents 98% precision and 63% recall. Class 2, which includes only the 4% of the test set, presents the lowest rates, i.e. 67% precision and 2% recall in tree classification and 45% precision and 4% recall in STZ. Overall, these findings indicate that both methods could predict adequately the Safety Tolerance Zone with high precision and slight differences.

Table 2. Results of STZ from tree classification with the best parameters determined

	0	**1**	**2**	**Total**
Accuracy	92%	96%	96%	92%
Precision	92%	98%	67%	92%
Recall	100%	63%	2%	92%
Specificity	48%	100%	100%	96%
G-means	48%	63%	2%	88%
F-means	96%	77%	4%	92%

**0 refers to normal phase, 1 refers to dangerous phase, 2 refers to avoidable accident phase*

Table 3. Results of STZ from SVC with the best parameters determined

	0	**1**	**2**	**Total**
Accuracy	92%	96%	96%	92%
Precision	92%	98%	45%	92%
Recall	100%	63%	4%	92%
Specificity	49%	100%	100%	96%
G-means	49%	63%	4%	88%
F-means	95%	77%	7%	92%

**0 refers to normal phase, 1 refers to dangerous phase, 2 refers to avoidable accident phase*

4.2. Prediction of time spent in each STZ level

Exploiting the predictions of the above models, the time spent in each STZ level was computed through the algorithm depicted in Figure 1. After examination of several alternatives, Support Vector Regression was selected and performed twice, utilizing the results of both STZ models. To evaluate the developed regression, the analyzed dataset was split to a 75%/25% training/test set ratio. Additionally, a hyperparameter tuning with 10-fold Cross-Validation was implemented, aiming to minimize the accuracy metric RMSE. The optimized hyperparameters C, epsilon and gamma were also manually tuned and the best performance combination is presented in Table 4.

Table 4. Results of SVR implementations with the best parameters determined

	Tree classification	**SVC**
train	0.8	0.8
kernel	radial	radial
gamma	0.65	0.65
epsilon	0.1	0.1
cost	25	25
MAE	250.85	251.09
MSE	276726	274938
RMSE	526.05	524.35
R^2	0.81	0.82

For the testing dataset, the obtained performance metrics were for the first SVM model: MAE = 250.85, MSE = 276726, RMSE=526.05 and R^2= 81.48%, while for the second one: MAE = 251.09, MSE = 274938, RMSE=524.35 and R^2= 81.60%. The values for R^2 denotes that the SVM algorithm predicts correctly the time drivers spent in each Safety Tolerance Zone for more than 80% of the data. Based on these metrics, SVM yields satisfactory results overall, but further models should be implemented to provide minimized MAE, MSE and RMSE metrics.

6. Discussion

In the present study, two separate analyses took place. First, two classification models, i.e. Conditional Inference Tree Classification and Support Vector Classification, were applied to predict the three phases of Safety Tolerance Zone for driving speed, i.e. Normal Driving, Danger and Avoidable Accident, based on naturalistic risk indicators. Subsequently, the predicted results were implemented in two versions of Support Vector Regression to predict how much time is spent in these levels, accounting the same risk indicators. This forms the main innovation of the present research, which is the machine learning analysis of real-time driving indicators in order to predict the duration of dangerous driving in terms of car speed.

With regards to the classification analysis, both performed methods provided accurate predictions with 92% accuracy, 96% specificity and 92% recall in total. Looking into regression results, Support Vector Machine models presented RMSE=526.0 and R^2= 81.48%, using STZ from Conditional Inference Tree Classification and RMSE=524.35 and R^2= 81.60%, using STZ from Support Vector Classification.

Nevertheless, there are some limitations and restrictions that should be mentioned. More specifically, the influence of socio-demographic characteristics or traffic environment was not taken into consideration in the present study. Based on the evidence that drivers react differently under different circumstances with respect to traffic conditions, it is of great interest to investigate average speed and duration combining traffic and questionnaire data.

The investigation of other significant factors could be also included in future research, such as drug abuse, alcohol consumption or the use of seat belt. As per further research directions, demographic characteristics such as gender, educational level, or driving experience could be also taken into account. Lastly, the experimental sample size could be strengthened in terms of size as well as in terms of country, region and transport mode (comparison among countries or different transport modes).

7. Conclusions

The current paper aims at providing a real-time prediction protocol for identifying the time that drivers spent in dangerous driving situations. By combining task complexity and coping capacity indicators, recorded from a naturalistic-driving experiment, state-of-the-art classifiers and regression approaches were exploited for providing results even with data collected at a 30-seconds aggregation level. Thus, data from 20 Belgian car drivers was utilized.

Towards that aim, the most reliable indicators of task complexity, such as time headway and distance travelled or weather conditions were assessed. In order to predict the time spent in each STZ level, two classification models, i.e. Conditional Inference Trees and Support Vector Machines, were developed. The obtained results were evaluated and then implemented to SVR models in order to predict the time drivers spent in each level.

Overall, the results of the study demonstrated that it is feasible not only to predict the safety level of each driver in real-time, but simultaneously predict how much time each driver is going to spend in each level. As the methodology presented in this paper is data-driven, researchers and practitioners, working in the field of road safety, could exploit the demonstrated results and methods using a large variety of naturalistic driving data, collected by a multitude of sensors to make roads much safer in the near future.

Within the i-DREAMS framework, the conclusions drawn from the current research serve as the base for building the mathematical models which are the backbone of the development of the i-DREAMS platform. Constructs to be measured are the driver's cognitive and behavioral state in terms of time spent in dangerous driving conditions as well as more stable characteristics which are known to impact safe driving and therefore, road safety. Another outcome will be a research database with rich information of simulator and on-road drives of hundreds of participants. Since the database aims to facilitate future research, it can be argued that if a vast amount of data on test subjects is obtained, research findings will be better validated. The testing, calibration and enhancement of the mathematical models during the i-DREAMS on-road and simulator experiments can assure not only an efficient and sufficient data analysis, but also timely initiation of the safety interventions.

Acknowledgements

This research was funded by the European Union's Horizon 2020 i-DREAMS project (Project Number: 814761) funded by European Commission under the MG-2-1-2018 Research and Innovation Action (RIA).

References

CardioID Technologies (2022). Heartmetrics Experts. Last assessed 15/05.22. Retrieved from: https://www.cardio-id.com/automotive/.

Das, A., Abdel-Aty, M., & Pande, A. (2009). Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. Journal of safety research, 40(4), 317-327.

Hothorn, T., Hornik, K., Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics. 15. 651-674. 10.1198/106186006X133933.

Hsu, Chih-wei & Chang, Chih-chung & Lin, Chih-Jen. (2003). A Practical Guide to Support Vector Classification.

Katrakazas, C., Michelaraki, E., Yannis, G., Kaiser, S., Brijs, K., Ross, V., ... & Taveira, R. (2020a). Technologies for safety interventions and assessment of their effectiveness.. Deliverable 2.2 of the EC H2020 project i-DREAMS.

Katrakazas, C., Michelaraki, E., Yannis, G., Kaiser, S., Senitschnig, N., Ross, V., ... & Taveira, R. (2020b). Toolbox of recommended data collection tools and monitoring methods and a conceptual definition of the Safety Tolerance Zone. Deliverable 3.2 of the EC H2020 project i-DREAMS.

Michelaraki, E., Katrakazas, C., Yannis, G., Filtness, A., Talbot, R., Hancox, G., ... & Taveira, R. (2021). Post-trip safety interventions: state-of-the-art, challenges, and practical implications. Journal of safety research, 77, 67-85.

Mobileye (2022). An Intel Company. Last assessed 15/05.22. Retrieved from: https://www.mobileye.com/

Sarda-Espinosa, A., Subbiah, S., & Bartz-Beielstein, T. (2017). Conditional inference trees for knowledge extraction from motor health condition data. Engineering Applications of Artificial Intelligence, 62, 26-37.

Vapnik, V. (1999). The nature of statistical learning theory. Springer science & business media.

Xue, Q., Wang, K., Lu, J. J., & Liu, Y. (2019). Rapid driving style recognition in car-following using machine learning and vehicle trajectory data. Journal of advanced transportation, 2019.

Yokoyama, D., & Toyoda, M. (2015). A large scale examination of vehicle recorder data to understand relationship between drivers' behaviors and their past driving histories. In 2015 IEEE International Conference on Big Data (Big Data) (pp. 2877-2879). IEEE.

Yu, H., & Kim, S. (2012). SVM Tutorial: Classification, Regression, and Ranking. Handbook of Natural Computing.