

This is the submitted manuscript of an article published by Taylor & Francis in Journal of Intelligent Transportation Systems on 08 Jul 2021, available online: <http://www.tandfonline.com/https://doi.org/10.1080/15472450.2021.1948412>

Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

**Please provide feedback**

Please support the ScholarWorks@UMBC repository by emailing [scholarworks-group@umbc.edu](mailto:scholarworks-group@umbc.edu) and telling us what having access to this work means to you and why it's important to you. Thank you.

## Driving Behavior Classification at Signalized Intersections Using Vehicle Kinematics: Application of Unsupervised Machine Learning

Nour O. Khanfar<sup>1</sup>, Mohammed Elhenawy<sup>2</sup>, Huthaifa I. Ashqar<sup>\*3</sup>, Qinaat Hussain<sup>4</sup>, and Wael K.M. Alhajyaseen<sup>4,5</sup>

<sup>1</sup>Arab American University, Natural, Engineering and Technology Sciences Department, P.O. Box 240, 13 Zababde, Jenin, Palestine.  
(E-mail: nour12@vt.edu)

<sup>2</sup>CARRS-Q, Centre for Accident Research and Road Safety, Queensland University of Technology, Queensland, Australia.  
(E-mail: mohammed.elhenawy@qut.edu.au)

<sup>3</sup>Precision Systems, Inc. Washington, DC, United States and University of Maryland Baltimore County, MD, United States.  
(E-mail: hiasqar@vt.edu)

<sup>4</sup>Qatar University, Qatar Transportation and Traffic Safety Centre, College of Engineering, P.O. Box 2713, Doha, Qatar.

<sup>5</sup>Qatar University, Department of Civil & Architectural Engineering, College of Engineering, P.O. Box 2713, Doha, Qatar.  
(E-mail: qinaat.hussain@qu.edu.qa, wyaseen@qu.edu.qa)

### ABSTRACT

Driving behavior is considered as a unique driving habit of each driver and has a significant impact on road safety. This study proposed a novel data-driven Machine Learning framework that can classify driving behavior at signalized intersections considering two different signal conditions. To the best of our knowledge, this is the first study that investigates driving behavior at signalized intersections with two different conditions that are mostly used in practice, i.e., the control setting with the signal order of green-yellow-red and a flashing green setting with the signal order of green-flashing green-yellow-red. A driving simulator dataset collected from participants at Qatar University's Qatar Transportation and Traffic Safety Center, driving through multiple signalized intersections, was used. The proposed framework extracts volatility measures from vehicle kinematic parameters including longitudinal speed and acceleration. K-means clustering algorithm with elbow method was used as an unsupervised machine learning to cluster driving behavior into three classes (i.e., conservative, normal, and aggressive) and investigate the impact of signal conditions. The framework confirmed that in general driving behavior at a signalized intersection reflects drivers' habits and personality rather than the signal condition, still, it manifests the intersection nature that usually requires drivers to be more vigilant and cautious. Nonetheless, the results suggested that flashing green condition could make drivers more conservative, which could be due to the limited capabilities of human to estimate the remaining distance and the prolonged duration of the additional flashing green interval. The proposed framework and findings of the study were promising that can be used for clustering drivers into different styles for different conditions and might be beneficial for policymakers, researchers, and engineers.

**Keywords:** Driving Behavior, Driving Style, Volatility Measures, Signalized intersection, Vehicle Kinematics.

### INTRODUCTION

Driving behavior is a reflection of a unique driving habit of drivers. It can be considered as a series of human activities which is reflected in drivers' actions and decisions during driving. Driving behavior has a significant impact on road safety. Since driving behavior is critical, it has attracted researchers to investigate several aspects using different methods. In general, identifying driving styles can be performed by two main types of studies: one is based on questionnaires. For example,

Sergio Useche *et al.* used a sample of 120 drivers to compare the incidence of physiological and representational indicators of stress and risky behaviors while driving [1]. Farah *et al.* used self-reported questionnaires to collect drivers' driving style indicators and to obtain driving behavior observations. However, questionnaire-based studies are subjective and may not be representative [2]. The second type of studies are based on the information of vehicle motion to recognize the driving style, which have emerged during the era of connected vehicles and the ability to collect big driving data. Feng *et al.* used real driving data of three human participants, which were collected using an instrumented vehicle, to classify driving style using the Support Vector Clustering approach [3].

In general, several algorithms and techniques may be used to recognize driving behavior. These algorithms and methods can be classified into two groups: methods based on machine learning, and methods based on rules. Aljaafreh *et al.* used fuzzy logic as a rule-based model to classify drivers as normal, aggressive, and very aggressive and recognized the driving events that fall into each of these categories [4]. The machine learning method is roughly divided into supervised and unsupervised learning approaches. For example, Ly *et al.* used a support vector machine (SVM) as a supervised method to explore the possibility of using the labeled vehicle's inertial sensors from the Controller Area Network (CAN) of a bus to build a profile of the drivers [5]. Assigning unknown data into categories by mining the underlying sources of unlabeled data is called the unsupervised machine learning method. For example, clustering and Principal Component Analysis (PCA) from exploratory statistics have been used to identify and explain drivers grouping according to their driving behavior [6]. However, the most used methods in driving style classification are the Support Vector Machine (SVM) [7], the Artificial Neural Network (ANN) [8][9], Random Forest Decision [10], and K-means [11].

In this paper, we used an unsupervised machine learning method to comprehensively cluster the different driving behavior in signalized roads into three styles: aggressive, normal, and conservative utilizing features extracted from vehicle kinematics. We used a dataset, collected by a driving simulator at Qatar University's Qatar Transportation and Traffic Safety Center. The driving simulator has the capability of collecting a wide range of data including speed, longitudinal acceleration, longitudinal position, number of accidents, number of red-light tickets, number of speeding tickets, pedal inputs, reaction time, and other information [12]. Furthermore, the used driving simulator was validated for different aspects of driving behavior: a) driving speed perception and actual speed [13]; b) geometric field of view [14]; and c) simulation sickness and the sense of presence [15]. Since most of the road crashes happened at road intersections [16], the simulator was designed to replicate real signalized intersection roads in Doha, Qatar. It is important to mention that Qatar is characterized with a very unique driver population where less than 10% of the population are Qatari Nationals [17], [18] while the rest are expatriates from different countries with various background cultures and habits [19], [20]. This indicates the challenges faced by road authorities to communicate with drivers, to enforce traffic rules, and to disseminate effective information to improve their driving behavior [21], [22]. Sixty-six drivers with a valid Qatari driver's license were invited to test different traffic signals conditions, each condition included two different situations based on the distance from the stop line on a road with a speed limit of 80 kph. The design of the study includes six intersections at 0.7, 1.77, 3.2, 6.1, 8.5, and 10.7 km from the starting point, respectively. In the first situation, drivers were in the indecision zone where there is a higher chance of red-light running. The second situation is that drivers likely were in the stopping zone when the signal turned yellow. In this case, drivers should stop to avoid red-light running. In this study, we only focused on two situations (i.e., indecision zone and likely-stopping zone) as there are some data missing for some drivers in the third situation (i.e., likely-go zone). However, this will be covered in a future study. This study used data from signalized intersections that are equipped with two traffic signal conditions that are widely used. The first one is the control condition, where the signal order is green, yellow, and then red. The other condition is called the flashing green condition, where the signal order is green, green-flashing, yellow, and then red. As shown in [12], at the onset of the yellow interval, the distance between the vehicle and stop line was 80 m and 95 m in the first and second situation, respectively. These

situations were proposed aiming that the first situation would be an indecision zone for a substantial number of drivers while the second one will be only for a limited number of (more aggressive) drivers. Various previous studies have used driving simulators data in evaluating the impact of human factors on road safety [23]–[27]. More information about the dataset can be found in [12]. Moreover, other studies have used driving simulators to collect driving data to be used for road safety investigation. For example, the Next Generation Simulation (NGSIM) dataset were used in many studies including [28], [29]. Wang *et al.* also used driving simulators to classify drivers into aggressive and normal. A driving simulator was also used to analyze driving behavior under rainy weather in [30]. In addition to driving simulators, one can collect data using smartphone sensors [31], which can be used to investigate dangerous driving recognition [32], mode recognition [33], [34], and other applications.

It is also worth offering definitions to the different driving behaviors, which can help in understanding the results of this study from driving through intersection perspective [35]–[38]. Conservative driver is defined as the one who drives in lower speed than the average operating speed of other vehicles, and the one who most likely will stop at an intersection during signal change indication (flashing green or flashing yellow); although he can safely clear the intersection (i.e., there is sufficient time to pass the intersection before red). Normal driver is defined as a driver who drives within the speed limit and cross the intersection during change interval without accelerating, if available time is sufficient, or stops if at his current speed cannot cross the intersection safely. However, aggressive driver is the one who should stop since the available time to cross at the onset of change interval (flashing green or yellow) based on their current speed is not sufficient, however they decide to accelerate and cross. Driver who makes a lane change near to an intersection or inside the intersection are also considered aggressive.

As the focus of this study is to explore driving behavior at intersection, it is very important to discuss drivers' indecisiveness at intersections in the literature. The (in)ability of drivers to decide whether to stop or go in the decision zone is mainly due to their limited human capabilities [39], [40]. Dilemma zone has been a research interest from early 1960s after it was first referred in literature [39]. There are two types of dilemma zone. Type I dilemma zone is deterministic and thus can be avoided by providing sufficient minimum yellow interval [39]. Another stochastic one, namely, Type II dilemma zone, which is also called Indecision zone [41], where it introduces more complexities for drivers attempting to perform decision-making process under mixed traffic conditions. There are many factors that affect drivers' indecisiveness at decision zone. For example, higher approach speeds and shorter distance to the stop line decreased the probability of stopping at the intersection [42]. Drivers are more likely be able to safely stop at the intersection if no flashing green exited at the intersection [43]. Shorter yellow duration, longer cycle length, and presence of pedestrians can make drivers more likely to stop at the intersections [44]. Gender and age of the drivers were found to have significant impact on the driver's decision [45]. Specifically, males were found to be more aggressive [46]. Flashing green ahead of the yellow signal tends to increase the number of early stops at the intersection [42].

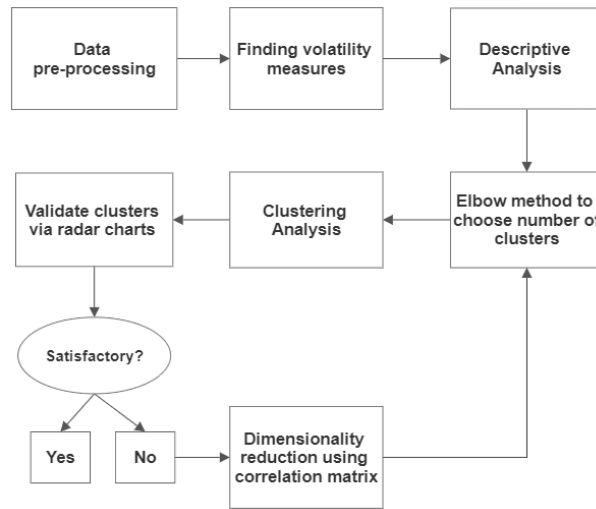
Driving behavior recognition has many useful applications. For example, driving style can be very useful in evaluating fuel consumption on the road [47]. Furthermore, driving assistant systems can be connected with the driving behavior recognition model to send an updated notification to the drivers to warn them that they are reached the aggressive driving style [48]. Moreover, insurance companies are interested to classify their driver's clients to offer the suitable premium accounts [49]. In addition to these applications, driving classification can be also used to allow autonomous vehicles to identify risky drivers and avoid them accordingly. Driving in urban environments in a mixed traffic can be very complicated and it is very challenging for autonomous driving systems to assess and distinguish neighboring human driven vehicles based on their risk taking to take proper actions and avoid severe conflicts. For the autonomous vehicles, replacing human drivers and enabling highly and fully automated driving systems, driving scenario understanding and drivers' behavior prediction are essential components. In deeply stochastic and uncertain traffic scenarios, comprehensive

prediction models intend not just to reproduce, but rather to encode the human driver behavior, which requires profound driving behavior understanding. That said, to the best of our knowledge, there is a notable gap in the literature in studying driving behavior at signalized intersections, which can be considered as one of the most hazardous locations in road networks [12]. Driving through intersections is relatively vulnerable due to several reasons including drivers' indecisiveness while approaching yellow interval at signalized intersection and the limited human capabilities to estimate the remaining time in yellow interval as well as the remaining distance to the stop line. This study addresses driving behavior in different road intersections with different traffic signal conditions. The study investigates the perception of aggressive or conservative driving styles and to what extent can vary in different conditions. The main contribution of this study could be concluded in two points. First, to develop a scalable data-driven machine learning framework to profile driving behavior at signalized intersections taking into consideration the perception of aggressive or conservative driving styles in terms of traffic signals conditions. Second, the study extracts several volatility measures from vehicle kinematics data and investigate their impact on the framework.

## **METHODS**

The main contribution of this study is to develop a data-driven machine learning framework to profile driving behavior at signalized intersections. In total, the design of the study includes six intersections with two different conditions and each condition. The first signal condition is the control condition which is the typical one when the signal order is green, yellow, and red. The second condition is the flashing green condition when the signal order is green, flashing green, yellow, and red. A more comprehensive details about the design of the conditions and the implementation of the indecision zone can be found here [12].

To classify driving behavior, the volatility measures were used as important safety parameters to distinguish each driving behavior, which have been used in many studies [7], [50]. Volatility measures can help in understanding driving decisions and actions. High driving volatility means high driving instability and thus more aggressive. We used descriptive analysis to thoroughly understand the extracted features and draw initial conclusions that can help in developing the framework. Afterwards, to prepare for clustering analysis, elbow method was used to determine the optimal number of clusters. The next step was using unsupervised machine learning to better deal with unlabeled data and cluster driving behavior employing extracted volatility measures as inputs. For this analysis, we used K-means as a clustering algorithm. To validate the suitability of the clusters, we used radar charts in which volatility measures are represented on axels that start from the center of the chart. The length of each bar depicts the value of each corresponding volatility measure in the cluster. At this point, if the results were considered reasonable, we stopped. If the results were not satisfactory, we used correlation matrix to reduce the number of features by removing measures with low correlation. Analysis can be repeated until the results are reasonable. This framework is illustrated in Figure 1. It is worth mentioning that as the data were collected from driving simulator, the effect of connectivity between vehicles such as collisions is not included within the scope of this model. The next section will describe the extraction of volatility features, and then K-means.



**Figure 1** Proposed Framework to Cluster Driving Behavior

### Volatility Measures

Previous studies developed different models using several volatility measures which have been extracted from various vehicle kinematics data including lateral acceleration [51], speed [50], vehicular jerk [50] and longitudinal acceleration [50], [52]. In this study, we utilized vehicle kinematics data, which can be collected from connected vehicles for example, to extract input features to the proposed framework. Table 1 shows volatility measures which was used as inputs in our framework.

**Table 1** Volatility Measures (where  $V$ : Speed,  $D_{long}$ : Longitudinal Deceleration,  $A_{long}$ : Longitudinal Acceleration, and  $AD_{long}$ : Longitudinal Deceleration or Acceleration)

Volatility Measure	Description	Equation
$DV_1$	Standard deviation of speed	$\sqrt{\frac{\sum_{i=1}^N (V_i - \bar{V})^2}{N}}$
$DV_2$	Standard deviation of longitudinal deceleration or acceleration	$\sqrt{\frac{\sum_{i=1}^N (AD_{long_i} - \bar{AD}_{long})^2}{N}}$
$DV_3$	Coefficient of variation of speed	$100 \times \frac{\sqrt{\frac{\sum_{i=1}^N (V_i - \bar{V})^2}{N}}}{\bar{V}}$
$DV_4$	Coefficient of variation of longitudinal acceleration	$100 \times \frac{\sqrt{\frac{\sum_{i=1}^N (A_{long_i} - \bar{A}_{long})^2}{N}}}{\bar{A}_{long}}$
$DV_5$	Coefficient of variation of longitudinal deceleration	$100 \times \frac{\sqrt{\frac{\sum_{i=1}^N (D_{long_i} - \bar{D}_{long})^2}{N}}}{\bar{D}_{long}}$
$DV_6$	Mean absolute deviation of speed	$\frac{\sum_{i=1}^N  V_i - \bar{V} }{N}$
$DV_7$	Mean absolute deviation of longitudinal acceleration	$\frac{\sum_{i=1}^N  A_{long_i} - \bar{A}_{long} }{N}$
$DV_8$	Quantile coefficient of variation of normalised speed	$100 \times \frac{Q_{V3} - Q_{V1}}{Q_{V3} + Q_{V1}}$ , where $Q_1$ and $Q_3$ are the sample 25 <sup>th</sup> and 75 <sup>th</sup> percentiles.
$DV_9$	Quantile coefficient of variation of longitudinal acceleration	$100 \times \frac{Q_{A_{long3}} - Q_{A_{long1}}}{Q_{A_{long3}} + Q_{A_{long1}}}$

$DV_{10}$	Quantile coefficient of variation of longitudinal deceleration	$100 \times \frac{Q_{D_{long}3} - Q_{D_{long}1}}{Q_{D_{long}3} + Q_{D_{long}1}}$
$DV_{11}$	Percentage of time the mean normalised speed exceeds the mean plus two standard deviations	$100 \times \frac{\sum_{i=1}^N (V_i \geq \bar{V} + 2\sigma)}{N}, \alpha = DV_1$
$DV_{12}$	Percentage of time the mean of longitudinal acceleration exceeds the mean plus two standard deviations	$100 \times \frac{\sum_{i=1}^N (A_{long_i} \geq \bar{A}_{long} + 2\sigma)}{N}, \alpha = DV_2$
$DV_{13}$	Percentage of time the mean longitudinal deceleration exceeds the mean plus two standard deviations	$100 \times \frac{\sum_{i=1}^N (D_{long_i} \geq \bar{D}_{long} + 2\sigma)}{N}, \alpha = DV_2$

### K-means Algorithm

K-means algorithm is an unsupervised machine learning technique. It aims to partition  $n$  observations into  $k$  clusters, in which each observation belongs to the cluster so that each group of observations clusters around the nearest centroid [49]. The outcome of the K-means algorithm is a group of clusters with corresponding centroids to minimize the following error function [53]:

$$E = \sum_{i=1}^k \sum_{x \in C_i} d(x, \mu(C_i))$$

where  $C_1, C_2, \dots, C_k$  are the number of clusters  $k$ ,  $\mu(C_i)$  is the centroid of the cluster  $C_i$ , and  $d(x, \mu(C_i))$  is the distance between the observation  $n$  and the centroid  $\mu(C_i)$ . In this study, Euclidean Distance is used. If  $x = \{x_1, x_2, \dots, x_n\}$  and  $\mu = \{\mu_1, \mu_2, \dots, \mu_n\}$  are the points and the clusters' centroids respectively, Euclidean Distance from  $x$  to  $\mu$  can be computed as follows:

$$d = \sqrt{\sum_{k=1}^n (x_k - \mu_k)^2}$$

Assuming  $D$  as the dataset, the steps to implement K-means can be described as follows:

*Step 1:* Randomly determined initial values for centroids ( $C_1, C_2, \dots, C_k$ ) from  $D$ .

*Step 2:* Use Euclidean Distance to assign observations to clusters that have nearest centroids ( $\min d$ ).

*Step 3:* Based on the outcome, recalculate the centroids of the new clusters based on the mean observations in each cluster.

*Step 4:* Reassign observations to the nearest new clusters based on the calculated centroids in *Step 3*.

*Step 5:* Repeat *Step 3* and *4* until no significant deviation in the error function.

*Step 6:* Report the outcomes.

K-means algorithm offers many advantages that makes it suitable for the proposed framework including it is relatively simple to implement and interpret its results, can be straightforwardly scaled to larger datasets, it guarantees convergence as the iterations proceed, and can be easily generalized to cluster different datasets and situations.

## RESULTS AND DISCUSSION

### Descriptive Analysis

Using the volatility measures described in Table 1, volatility measures for each participant were found using the observed features at each of the six intersections. Table 2 shows the descriptive summary statistics of the volatility measures for each condition. According to the results, driving volatility measures have relatively similar values for control condition compared to flashing green condition. For example, the mean values for the thirteen volatility measures are very close when compared between the two different conditions. This initially indicates that there is no significant difference in driving behavior between the two conditions. This could be expected as drivers usually tend to exercise more caution (i.e., be relatively more alarmed) when approaching

a signalized intersection. Specifically, experienced drivers usually develop the habit of slowing down as they approach to a traffic light to avoid running in red or safely clear the intersection.

**Table 2** Descriptive Summary Statistics of the Volatility Measures for Each Condition

Volatility Measures	Control Condition				Flashing Green Condition			
	Max	Min	Mean	SD	Max	Min	Mean	SD
$DV_1$	12.26	0.01	7.45	3.39	11.85	0.07	7.27	3.39
$DV_2$	2.56	0.00	1.53	0.60	2.99	0.00	1.52	0.62
$DV_3$	144.02	0.07	66.22	34.03	114.89	0.30	64.35	33.80
$DV_4$	210.66	16.40	87.34	33.72	221.70	0.00	89.17	34.94
$DV_5$	0.00	-199.30	-98.72	36.80	-18.07	-223.55	-105.96	36.03
$DV_6$	11.26	0.01	6.91	3.23	10.98	0.06	6.76	3.23
$DV_7$	1.78	0.00	0.96	0.36	2.09	0.00	0.93	0.36
$DV_8$	100.00	0.07	73.59	38.78	100.00	0.27	73.09	39.05
$DV_9$	99.46	6.37	65.91	24.47	98.70	0.00	66.04	25.00
$DV_{10}$	0.00	-99.54	-65.17	24.76	-5.15	-99.57	-70.23	23.40
$DV_{11}$	31.76	0.00	0.12	1.71	22.29	0.00	0.13	1.56
$DV_{12}$	98.53	0.00	26.46	20.93	91.44	0.00	24.58	19.26
$DV_{13}$	100.00	0.00	23.11	18.24	94.48	0.00	21.73	16.54

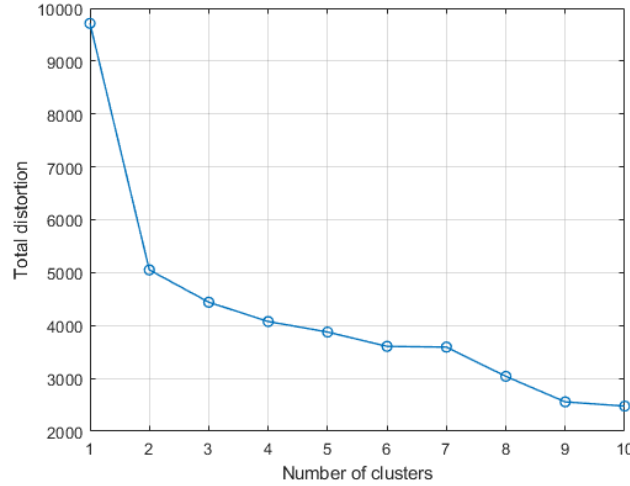
### Clustering Analysis

To prepare the data for clustering, we standardized data to make classification of each observation analogous across different features. Observations in the dataset denotes a driving behavior event across the two different conditions. Each observation will be clustered using K-means. In cluster analysis, we used *elbow method*, which is a heuristic used in determining the number of clusters in a dataset. The method consists of plotting the explained variation as a function of the number of clusters and choosing the elbow of the curve as the optimal number of clusters ( $k$ ) to use [54]. We used MathWorks's MATLAB software to implement K-means clustering algorithm [55]. Before we started with clustering driving behavior on the intersection level, we investigated K-means clustering on the condition level. As we note in the **Descriptive Analysis** section, clusters were found to be uncorrelated with the two different conditions at the signalized intersections. That said, we shall continue our clustering analysis ignoring the signal condition for the time being. We consider the signal conditions at the intersections and their impact on the cluster analysis once the framework is satisfactory and validated, as reported in **Clustering Analysis of Signal Conditions** section.

We used the elbow method to find the optimal number of clusters for K-means algorithm using the thirteen volatility measures. As Figure 2 shows, the optimal number of clusters can be chosen as two or three clusters as it produces a relatively low total distortion and can be physically interpreted. We tested both cases to compare their results. Each cluster was labeled as 1, 2, or 3 and each of them indicates a classified driving behavior. Determining which driving behavior is assigned to a cluster was based on the mean values of classification features. After performing K-means using all possible features for  $k = 2$  and  $k = 3$ , results were represented in Table 3 and Figure 3. Table 3 shows the scaled cluster centers using all possible features (i.e., 13 volatility measures) with  $k = 3$  and  $k = 2$ . Figure 3 depicts radar charts for the same cases. Volatility measures are represented on axels that start from the center of the chart. The length of each bar depicts the value of each corresponding volatility measure in the cluster. Connecting the spokes using a line creates regions. The larger region in size represents the higher measures' values in



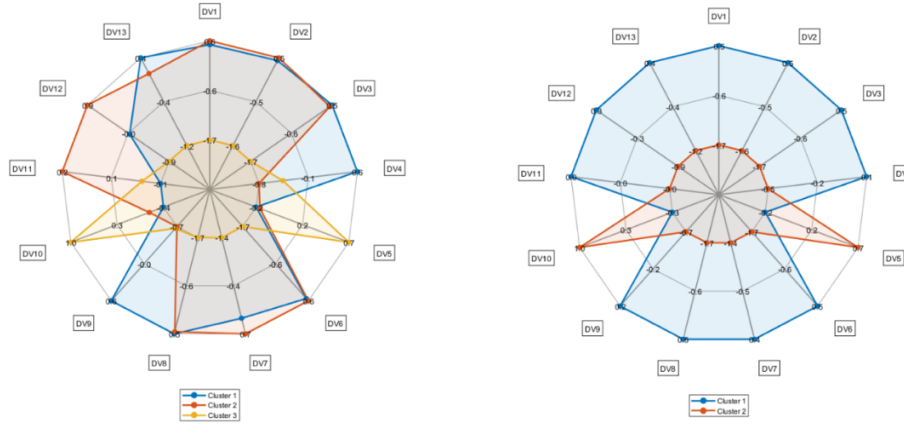
the cluster. Aggressive and risky driving behavior usually has high values of volatilities [49], [51]. When  $k = 3$ , we will assign the large, medium, and small regions (i.e., Clusters 1, 2, and 3) to aggressive, normal, and conservative driving behaviors, respectively. On the other hand, when  $k = 2$ , we will assign the large and small regions (i.e., Clusters 1 and 2) to normal and conservative driving behaviors, respectively.



**Figure 2** Visualization of Elbow Method for K-means Using All Possible Volatility Measures

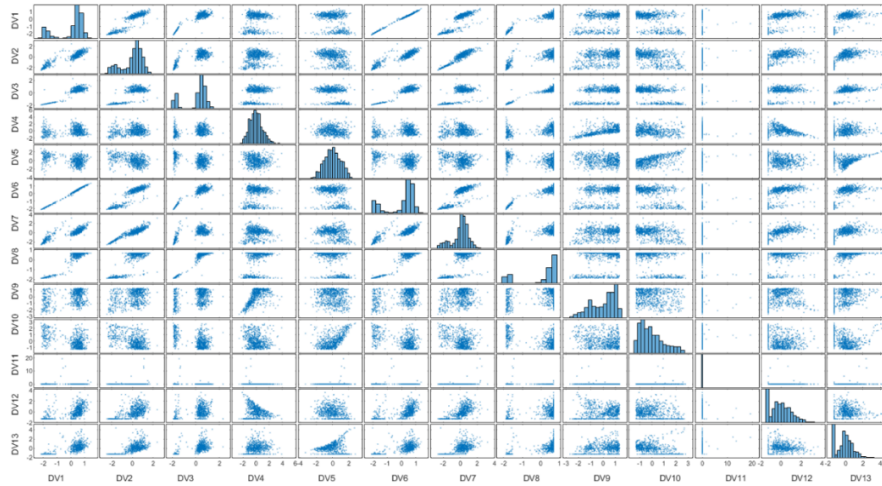
**Table 3** The Scaled Cluster Centers Using All Possible Features with  $k = 3$  and  $k = 2$

Volatility Measures	With All Possible Features				
	$k = 3$			$k = 2$	
	Cluster 1 (Conservative)	Cluster 2 (Normal)	Cluster 3 (Aggressive)	Cluster 1 (Conservative)	Cluster 2 (Normal)
$DV_1$	-1.73	0.48	0.61	-1.73	0.52
$DV_2$	-1.56	0.44	0.56	-1.56	0.48
$DV_3$	-1.71	0.53	0.49	-1.71	0.51
$DV_4$	-0.45	0.54	-0.76	-0.45	0.13
$DV_5$	0.71	-0.27	-0.10	0.71	-0.22
$DV_6$	-1.73	0.48	0.60	-1.73	0.52
$DV_7$	-1.42	0.30	0.73	-1.42	0.44
$DV_8$	-1.75	0.54	0.50	-1.75	0.52
$DV_9$	-0.67	0.57	-0.62	-0.67	0.20
$DV_{10}$	0.96	-0.37	-0.11	0.96	-0.29
$DV_{11}$	-0.03	-0.08	0.19	-0.03	0.01
$DV_{12}$	-0.91	-0.08	1.05	-0.91	0.27
$DV_{13}$	-1.19	0.42	0.20	-1.19	0.35
Sample Size	18	42	3	18	45



**Figure 3** Radar Charts of the Resulted Clusters Using All Possible Features with  $k = 3$  and  $k = 2$

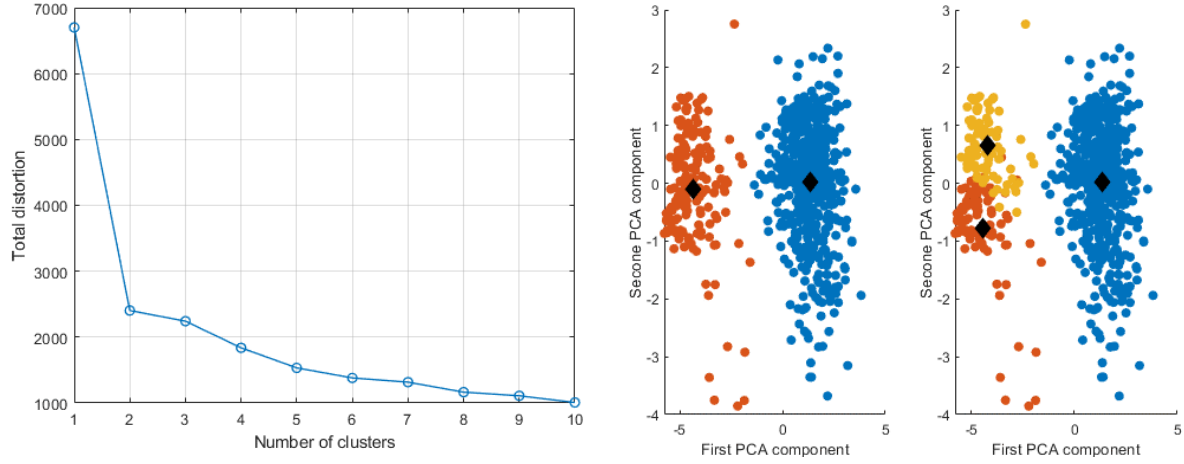
We used radar charts to validate the clusters for both  $k = 3$  and  $k = 2$  using all possible features. As shown in Figure 3, radar charts shows that there are some features' centroids that has lower/higher values in all dimensions, which made the regions that represent clusters in the radar charts rather distorted. This unstable result means that there might be some indiscriminatory features that would be considered as noise in the input features and might weaken the framework. To investigate this, we visually inspected the correlation matrix of the features, as shown in Figure 4. We created a matrix of plots showing the correlations among pairs of volatility measures using Pearson's linear correlation coefficient. Histograms of the variables appears along the matrix diagonal; scatter plots of variable pairs appear in the off diagonal. Figure 4 clearly shows that  $DV_4$ ,  $DV_5$ ,  $DV_{10}$ , and  $DV_{11}$  are noisy and no trend can be detected and thus we reduced the dimensions of the input features by excluding them.



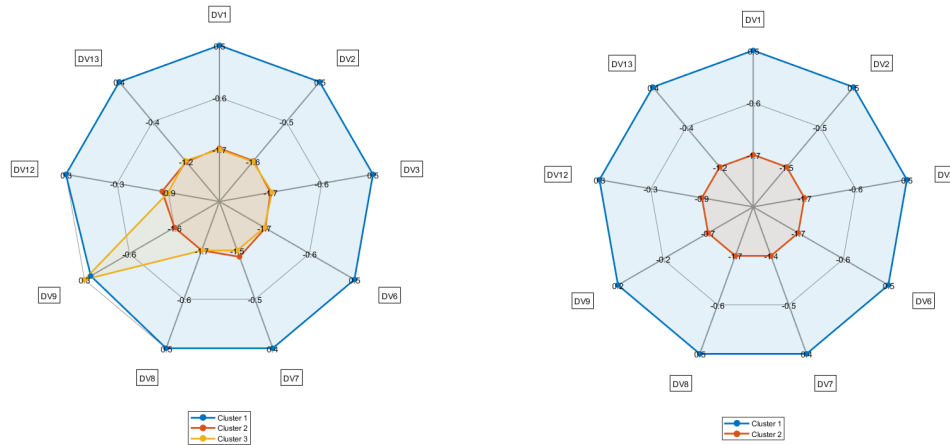
**Figure 4** Correlation Matrix for All Possible Features

Subsequently, we used the elbow method to find the optimal number of clusters using the remaining nine volatility measures. As Figure 5 (left) depicts, the optimal number of clusters is still the same, which can be two or three clusters. Therefore, we tested both cases to compare their results. We used K-means to cluster driving behavior  $k = 3$  and  $k = 2$ . We used Principal Component Analysis (PCA) to visualize the clusters in 2-dimentional space, as shown in Figure 5 (right). To validate the clusters, we created the radar charts for  $k = 3$  and  $k = 2$ , as shown in Figure 6. Moreover, Table 4 shows the scaled cluster centers using all possible features (i.e., 13 volatility measures) with  $k = 3$  and  $k = 2$ . Results showed that after decreasing the input features, the clusters are more stable, especially when  $k = 2$ . That said, participants can be

classified into conservative and normal driving behaviors. The cluster for normal driving behavior is more congested, which means more participants can be classified in this cluster. This note is also consistent with the results presented in Table 4. Once again, it is also possible to classify participants to three clusters (i.e., conservative, normal, and aggressive), but we argue evidence that supports two clusters are stronger.



**Figure 5** Elbow Method Using Nine Volatility Measures (Left), and Representation of Results in 2-Dimensional Space using PCA (Right).



**Figure 6** Radar Charts of the Resulted Clusters Using Nine Features with  $k = 3$  and  $k = 2$

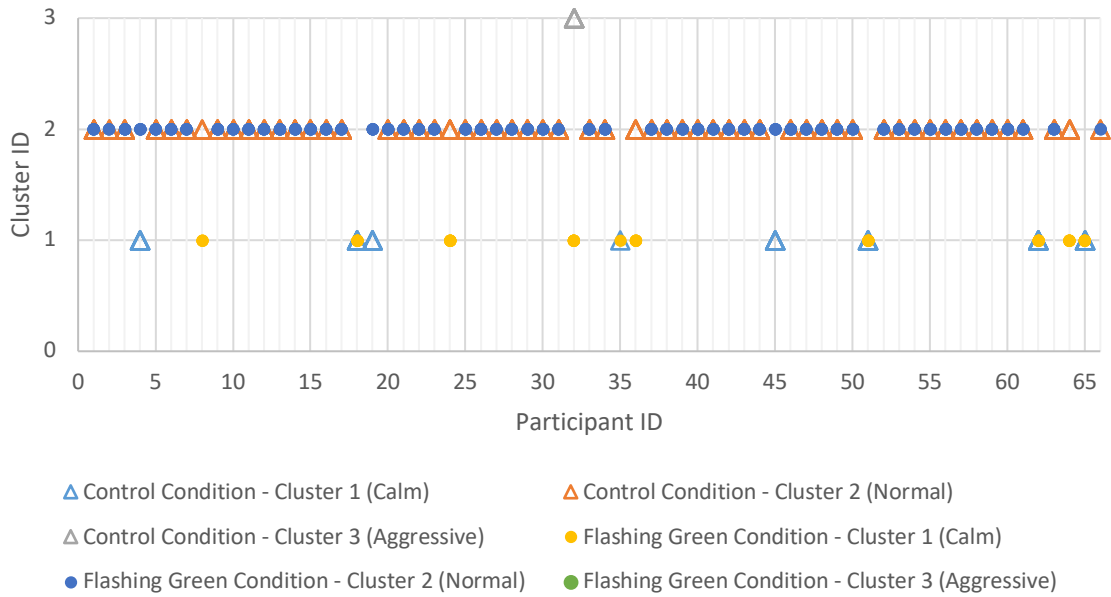
**Table 4** The Scaled Cluster Centers Using Nine Features with  $k = 3$  and  $k = 2$

Volatility Measures	With Nine Features				
	$k = 3$			$k = 2$	
	Cluster 1 (Conservative)	Cluster 2 (Normal)	Cluster 3 (Aggressive)	Cluster 1 (Conservative)	Cluster 2 (Normal)
$DV_1$	-1.80	0.54	-1.13	-1.73	0.52
$DV_2$	-1.72	0.47	-0.30	-1.55	0.48
$DV_3$	-1.74	0.54	-1.43	-1.71	0.52
$DV_4$	-	-	-	-	-
$DV_5$	-	-	-	-	-
$DV_6$	-1.79	0.54	-1.19	-1.73	0.52

$DV_7$	-1.65	0.42	0.24	-1.41	0.44
$DV_8$	-1.77	0.56	-1.54	-1.75	0.53
$DV_9$	-0.62	0.21	-0.85	-0.67	0.20
$DV_{10}$	-	-	-	-	-
$DV_{11}$	-	-	-	-	-
$DV_{12}$	-1.11	0.28	0.03	-0.91	0.27
$DV_{13}$	-1.22	0.34	-0.40	-1.18	0.35
Sample Size	13	52	1	15	51

### Clustering Analysis of Signal Conditions

On one hand, an important purpose of this study is to develop a framework that will be able to profile driving behavior traveling through signalized intersections. On the other hand, the study aims to investigate whether the perception of different driving behavior varies in two traffic signal conditions including the control and flashing green conditions. We noted in the **Descriptive Analysis** section that driving behavior was initially found to be uncorrelated with the two different conditions at the signalized intersections. We used our proposed framework to cluster driving behavior taking into consideration the traffic signal condition at the intersections. We used the nine features that we argued they resulted in more stable clusters including standard deviation of speed and longitudinal deceleration or acceleration, coefficient of variation of speed, mean absolute deviation of speed and longitudinal acceleration, quantile coefficient of variation of normalised speed and longitudinal acceleration, percentage of time when the mean of longitudinal acceleration and deceleration exceeds the mean plus two standard deviations. In our dataset, there are 61 participants who drove through six intersections twice, and another five participants who drove the same intersections only once. We investigated the impact of traffic signal condition (i.e., control and flashing green conditions) on driving behavior for participants using  $k = 3$  in K-means. Figure 7 depicts the results of classifying the participants for the two conditions. As we argued before, it seems that most of the drivers (58 out of 66) have the same driving behavior while driving through a signalized intersection regardless of the traffic signal condition. It is also notable that most of the drivers have normal driving behavior, drivers who have conservative driving behavior comes next, and finally, only one driver has an aggressive driving behavior while going through intersections that were treated with control condition. However, it appears that in five out of the eight cases, where the behavior differs in terms of conditions, flashing green condition has changed the driving behavior from normal or aggressive to conservative. These results indicate that overall driving behavior on a signalized intersection reflects its nature, which requires drivers to be alarmed if stopping might be needed and cautious of other vehicles. Moreover, driving behavior of participants seems to mainly stem from their habits and personality rather than signal condition of the intersection. Results also showed that flashing green condition makes drivers plan ahead of time and thus avoid sharp acceleration and deceleration more frequently. Although flashing green induces more conservative driving behavior, as it encourages drivers to stop at intersections, it increases the probability of inconsistent stopping behavior which could lead to rear-end collisions. Higher variations in drivers' speeds and decelerations in flashing green scenario could increase the indecision zone due to the two extra seconds (flashing green) provided to the drivers before the yellow interval [12].



**Figure 7** Results of Classifying Participants for Two Control Conditions

## CONCLUSION

Based on a driving simulator dataset of Qatar Transportation and Traffic Safety Center, this study developed a scalable framework to classify driving behavior at signalized intersection considering two different signal conditions including control and flashing green conditions. This study applied K-means as an unsupervised machine learning to classify driving behavior using features extracted from volatility measures of vehicle kinematics. Results of the proposed framework were promising and might be beneficial for policymakers, researchers, and engineers. We found that same driving behavior can be detected while driving through a signalized intersection regardless of the traffic signal condition. Still, there is some evidence to suggest that flashing green condition can change driving behavior from normal or aggressive to conservative. Although flashing green induces more conservative driving behavior, as it encourages drivers to stop at intersections even if they could cross the intersections safely, it increases the probability of inconsistent stopping behavior which could lead to rear-end collisions. The overall driving behavior at a signalized intersection reflects drivers' habits and personality rather than the signal condition, still, it manifests the intersection nature that usually requires more cautious. Nonetheless, there are several future recommendations to leverage this framework. First, increasing the number of participants will enrich the study and could be representable enough to investigate aggressive driving behavior. Second, different dataset from other simulators or geographical locations could further leverage this study to increase its veracity. Third, future work can attempt to calibrate driving behaviors with car-following models and then classify the drivers considering the parameters of the resulted car-following models. In that sense, the output can be contributed to commonly applied simulators. Finally, investigating demographic information for participants could be beneficial in deeply understanding the impact of personality on driving behavior. Still, the result of this study primarily contains significant finding in this field to build on, the first of its kind, and sheds the light on significant new understanding on driving behavior at signalized intersections.

## REFERENCES

- [1] S. Useche, A. Serge, and F. Alonso, "Risky behaviors and stress indicators between novice and experienced drivers," *American Journal of Applied Psychology*, vol. 3, no. 1, pp. 11–14, 2015.
- [2] Y. Peng, L. Cheng, Y. Jiang, and S. Zhu, "Examining Bayesian network modeling in identification of dangerous driving behavior," *PLoS one*, vol. 16, no. 8, p. e0252484, 2021.

- [3] Y. Feng, S. Pickering, E. Chappell, P. Iravani, and C. Brace, “A support vector clustering based approach for driving style classification,” *International Journal of Machine Learning and Computing*, vol. 9, no. 3, pp. 344–350, 2019.
- [4] A. Aljaafreh, N. Alshabat, and M. S. N. Al-Din, “Driving style recognition using fuzzy logic,” in *2012 IEEE International Conference on Vehicular Electronics and Safety (ICVES 2012)*, 2012, pp. 460–463.
- [5] M. van Ly, S. Martin, and M. M. Trivedi, “Driver classification and driving style recognition using inertial sensors,” in *2013 IEEE Intelligent Vehicles Symposium (IV)*, 2013, pp. 1040–1045.
- [6] Z. Constantinescu, C. Marinoiu, and M. Vladoiu, “Driving style analysis using data mining techniques,” *International Journal of Computers Communications & Control*, vol. 5, no. 5, pp. 654–663, 2010.
- [7] W. Wang, J. Xi, A. Chong, and L. Li, “Driving style classification using a semisupervised support vector machine,” *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 5, pp. 650–660, 2017.
- [8] C. MacAdam, Z. Bareket, P. Fancher, and R. Ervin, “Using neural networks to identify driving style and headway control behavior of drivers,” *Vehicle System Dynamics*, vol. 29, no. S1, pp. 143–160, 1998.
- [9] P. Brombacher, J. Masino, M. Frey, and F. Gauterin, “Driving event detection and driving style classification using artificial neural networks,” in *2017 IEEE International Conference on Industrial Technology (ICIT)*, 2017, pp. 997–1002.
- [10] J. Xie and M. Zhu, “Maneuver-based driving behavior classification based on random forest,” *IEEE Sensors Letters*, vol. 3, no. 11, pp. 1–4, 2019.
- [11] W. Wang and J. Xi, “A rapid pattern-recognition method for driving styles using clustering-based support vector machines,” in *2016 American Control Conference (ACC)*, 2016, pp. 5270–5275.
- [12] Q. Hussain, W. K. M. Alhajyaseen, K. Brijs, A. Pirdavani, and T. Brijs, “Innovative countermeasures for red light running prevention at signalized intersections: A driving simulator study,” *Accident Analysis & Prevention*, vol. 134, p. 105349, 2020.
- [13] Q. Hussain, W. K. M. Alhajyaseen, A. Pirdavani, N. Reinolsmann, K. Brijs, and T. Brijs, “Speed perception and actual speed in a driving simulator and real-world: A validation study,” *Transportation research part F: traffic psychology and behaviour*, vol. 62, pp. 637–650, 2019.
- [14] Q. Hussain, M. Almallah, W. K. M. Alhajyaseen, and C. Dias, “Impact of the geometric field of view on drivers’ speed perception and lateral position in driving simulators,” *Procedia Computer Science*, vol. 170, pp. 18–25, 2020.
- [15] M. Almallah, Q. Hussain, N. Reinolsmann, and W. K. M. Alhajyaseen, “Driving simulation sickness and the sense of presence: Correlation and contributing factors,” *Transportation research part F: traffic psychology and behaviour*, vol. 78, pp. 180–193, 2021.
- [16] H. Huang, D. Wang, L. Zheng, and X. Li, “Evaluating time-reminder strategies before amber: common signal, green flashing and green countdown,” *Accident Analysis & Prevention*, vol. 71, pp. 248–260, 2014.
- [17] C. Timmermans *et al.*, “Analysis of road traffic crashes in the State of Qatar,” *Int J Inj Contr Saf Promot*, vol. 26, no. 3, pp. 242–250, 2019.
- [18] A. Soliman, W. Alhajyaseen, R. Alfar, and I. Alkaabi, “Changes in driving behavior across age cohorts in an Arab culture: the case of state of Qatar,” *Procedia Comput Sci*, vol. 130, pp. 652–659, 2018.
- [19] C. Timmermans, W. Alhajyaseen, N. Reinolsmann, H. Nakamura, and K. Suzuki, “Traffic safety culture of professional drivers in the State of Qatar,” *IATSS research*, vol. 43, no. 4, pp. 286–296, 2019.
- [20] C. P. M. Timmermans, W. K. M. Alhajyaseen, V. Ross, and H. Nakamura, “Introducing a multi-variate classification method: Risky driving acceptance among different heterogeneous driver sub-cultures,” *J Safety Res*, vol. 73, pp. 81–91, 2020.
- [21] A. Toriumi *et al.*, “A multi-country survey for collecting and analyzing facts related to road traffic safety: legislation, enforcement, and education for safer drivers,” *IATSS research*, vol. 46, no. 1, pp. 14–25, 2022.
- [22] K. Suzuki, K. Tang, W. Alhajyaseen, K. Suzuki, and H. Nakamura, “An international comparative study on driving attitudes and behaviors based on questionnaire surveys,” *IATSS research*, vol. 46, no. 1, pp. 26–35, 2022.
- [23] D. Llopis-Castelló, F. J. Camacho-Torregrosa, J. Marín-Morales, A. M. Pérez-Zuriaga, A. García, and J. F. Dols, “Validation of a low-cost driving simulator based on continuous speed profiles,” *Transp Res Rec*, vol. 2602, no. 1, pp. 104–114, 2016.

- [24] L. N. Boyle and J. D. Lee, "Using driving simulators to assess driving safety," *Accident Analysis & Prevention*, vol. 42, no. 3, Elsevier, pp. 785–787, 2010.
- [25] Y. Ali, A. Sharma, M. M. Haque, Z. Zheng, and M. Saifuzzaman, "The impact of the connected environment on driving behavior and safety: A driving simulator study," *Accident Analysis & Prevention*, vol. 144, p. 105643, 2020.
- [26] X. Yan, M. Abdel-Aty, E. Radwan, X. Wang, and P. Chilakapati, "Validating a driving simulator using surrogate safety measures," *Accident Analysis & Prevention*, vol. 40, no. 1, pp. 274–288, 2008.
- [27] T. Horberry, J. Anderson, and M. A. Regan, "The possible safety benefits of enhanced road markings: a driving simulator evaluation," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 9, no. 1, pp. 77–87, 2006.
- [28] N. Raju, S. Arkatkar, S. Easa, and G. Joshi, "Developing extended trajectory database for heterogeneous traffic like NGSIM database," *Transportation Letters*, pp. 1–10, 2021.
- [29] S. H. Hamdar and H. S. Mahmassani, "Driver car-following behavior: From discrete event process to continuous set of episodes," 2008.
- [30] Alasan Jarju and Mansoureh Jeyhani, "Drivers' Behavior Analysis under Reduced Visibility Condition Using a Driving Simulator," *Journal of Traffic and Logistics Engineering*, vol. 6, Dec. 2018.
- [31] H. I. Ashqar, M. H. Almannaa, M. Elhenawy, H. A. Rakha, and L. House, "Smartphone transportation mode recognition using a hierarchical machine learning classifier and pooled features from time and frequency domains," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 244–252, 2018.
- [32] D. A. Johnson and M. M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2011, pp. 1609–1615.
- [33] H. I. Ashqar, M. Elhenawy, M. Masoud, A. Rakotonirainy, and H. A. Rakha, "Vulnerable road user detection using smartphone sensors and recurrence quantification analysis," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 1054–1059.
- [34] M. Elhenawy, H. I. Ashqar, M. Masoud, M. H. Almannaa, A. Rakotonirainy, and H. A. Rakha, "Deep transfer learning for vulnerable road users detection using smartphone sensors data," *Remote Sensing*, vol. 12, no. 21, p. 3508, 2020.
- [35] W. Vanlaar, H. Simpson, D. Mayhew, and R. Robertson, "Aggressive driving: A survey of attitudes, opinions and behaviors," *J Safety Res*, vol. 39, no. 4, pp. 375–381, 2008.
- [36] D. E. Miles and G. L. Johnson, "Aggressive driving behaviors: are there psychological and attitudinal predictors?," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 6, no. 2, pp. 147–161, 2003.
- [37] L. Tasca, *A review of the literature on aggressive driving research*. Citeseer, 2000.
- [38] D. Shinar, "Aggressive driving: the contribution of the drivers and the situation," *Transportation Research Part F: traffic psychology and behaviour*, vol. 1, no. 2, pp. 137–160, 1998.
- [39] D. Gazis, R. Herman, and A. Maradudin, "The problem of the amber signal light in traffic flow," *Oper Res*, vol. 8, no. 1, pp. 112–132, 1960.
- [40] B. K. Pathivada and V. Perumal, "Analyzing dilemma driver behavior at signalized intersection under mixed traffic conditions," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 60, pp. 111–120, 2019, doi: <https://doi.org/10.1016/j.trf.2018.10.010>.
- [41] S. G. Machiani and M. Abbas, "Safety surrogate histograms (SSH): A novel real-time safety assessment of dilemma zone related conflicts at signalized intersections," *Accident Analysis & Prevention*, vol. 96, pp. 361–370, 2016.
- [42] H. Köll, M. Bader, and K. W. Axhausen, "Driver behaviour during flashing green before amber: a comparative study," *Accident Analysis & Prevention*, vol. 36, no. 2, pp. 273–280, 2004.
- [43] K. Long, Y. Liu, and L. D. Han, "Impact of countdown timer on driving maneuvers after the yellow onset at signalized intersections: An empirical study in Changsha, China," *Saf Sci*, vol. 54, pp. 8–16, 2013.
- [44] T. J. Gates, D. A. Noyce, L. Laracuente, and E. v Nordheim, "Analysis of dilemma zone driver behavior at signalized intersections," *Transportation Research Record*, vol. 2030, pp. 29–39, 2007.
- [45] H. Rakha, A. Amer, and I. El-Shawarby, "Modeling driver behavior within a signalized intersection approach decision–dilemma zone," *Transportation Research Record*, vol. 2069, no. 1, pp. 16–25, 2008.
- [46] G.-L. Chang, M. Franz, and Y. Lu, "Design and Evaluation of a dynamic dilemma zone system for a high speed rural intersection," Maryland. State Highway Administration, 2012.

- [47] A. Alessandrini, A. Cattivera, F. Filippi, and F. Ortenzi, “Driving style influence on car CO2 emissions,” 2012.
- [48] G. A. M. Meiring and H. C. Myburgh, “A review of intelligent driving style analysis systems and related artificial intelligence algorithms,” *Sensors*, vol. 15, no. 12, pp. 30653–30682, 2015.
- [49] A. Mohammadnazar, R. Arvin, and A. J. Khattak, “Classifying travelers’ driving style using basic safety messages generated by connected vehicles: Application of unsupervised machine learning,” *Transportation research part C: emerging technologies*, vol. 122, p. 102917, 2021.
- [50] M. Kamrani, R. Arvin, and A. J. Khattak, “Extracting useful information from Basic Safety Message Data: an empirical study of driving volatility measures and crash frequency at intersections,” *Transp Res Rec*, vol. 2672, no. 38, pp. 290–301, 2018.
- [51] R. Arvin, M. Kamrani, and A. J. Khattak, “How instantaneous driving behavior contributes to crashes at intersections: extracting useful information from connected vehicle message data,” *Accident Analysis & Prevention*, vol. 127, pp. 118–133, 2019.
- [52] R. Arvin, M. Kamrani, and A. J. Khattak, “The role of pre-crash driving instability in contributing to crash intensity using naturalistic driving data,” *Accident Analysis & Prevention*, vol. 132, p. 105226, 2019.
- [53] G. Gan, C. Ma, and J. Wu, *Data clustering: theory, algorithms, and applications*. SIAM, 2020.
- [54] D. Marutho, S. H. Handaka, and E. Wijaya, “The determination of cluster number at k-mean using elbow method and purity evaluation on headline news,” in *2018 International Seminar on Application for Technology of Information and Communication*, 2018, pp. 533–538.
- [55] MathWorks - MATLAB, “K-means Clustering.”  
<https://www.mathworks.com/help/stats/kmeans.html>.