Canadian Journal of Civil Engineering

Human-like Speed Modelling for Autonomous Vehicles during Car-Following at Intersections

Journal:	<i>Canadian Journal of Civil Engineering</i>
Manuscript ID	cjce-2020-0761.R1
Manuscript Type:	Article
Date Submitted by the Author:	08-Feb-2021
Complete List of Authors:	Tawfeek, Mostafa; Ain Shams University, Public Works Department
Keyword:	Signalized intersections, Unsignalized intersections, Naturalistic driving, Driver behavior, Autonomous vehicles
Is the invited manuscript for consideration in a Special Issue? :	Not applicable (regular submission)

SCHOLARONE™
Manuscripts

Human-like Speed Modelling for Autonomous Vehicles during Car-Following at Intersections

Mostafa H. Tawfeek, Ph.D.^a

^a Public Works Department, Ain Shams University, Cairo, Egypt

Email: mostafa.h.tawfeek@eng.asu.edu.eg, ORCID: 0000-0002-6461-7187

ABSTRACT

This study aims at modelling drivers' speed in car-following during braking situations at intersections to estimate a safe comfortable human-like speed at the minimum car-following distance for Autonomous Vehicles (AV). Several car-following behavioral measures at different times before reaching the minimum following distance and the intersection control type (signalized or unsignalized) were extracted to train the model using three machine learning techniques. The results showed that the XGBoost model is superior when compared to other techniques with R-squared values of 0.99 and 0.97 for training and testing datasets. The results also indicated that the control type impact driver speed at the minimum following distance. The modelled speed will provide more comfortable speed to the AV riders and will not violate the expectations of the surrounding traditional vehicle drivers. Also, the proposed model can be adopted to enhance current car-following models by considering the effect of intersections and its control type.

Keywords: Signalized intersections, Unsignalized intersections, Principle Component Analysis, XGBoost, Neural networks, Naturalistic driving, Driver behavior, Autonomous vehicles, Car-following

INTRODUCTION

Autonomous vehicles (AV) technologies are expected to proactively prevent collisions by alleviating driver error across various automation levels. However, it is expected that the implementation of the AV technologies will be gradual over a long transition period until reaching high market penetration rates of fully AV. Therefore, it is essential to ensure safe interactions between the traditional (i.e., human-driven) vehicles and AV during this transitional period. These safe interactions can be attained when the AV imitates a safe human driving behavior, which will not violate the human drivers' expectations. Also, the replication of human driving behavior in AV will enhance drivers' comfort and trust in the AV (Wang et al. 2016; Zhu et al. 2018).

Several studies have been conducted to model, mimic, and integrate human-like car-following behavior for AV to ensure drivers' safety and comfort. To facilitate this integration, some specific situations of car-following were considered while developing these models such as cut-in behavior during car-following (Fu et al. 2019), distracted driving (Przybyla et al. 2012), human perception limitations (Bevrani and Chung 2012), and car-following during braking (Tawfeek and El-Basyouny 2018b). Although the surrounding environment significantly affects the car-following behavior (Saifuzzaman and Zheng 2014; Tawfeek and El-Basyouny 2020), the impact of surrounding infrastructure (e.g., intersections) was only considered in limited contexts. For example, the impact of green signal countdown devices on car-following modeling at signalized intersections was investigated (Yu and Shi 2015). Also, Tang et al. (Tang et al. 2018) formulated a speed guidance strategy based on car-following to minimize fuel consumption. The proposed strategy was validated in simulation using hypothetical scenarios. However, the proposed strategy did not account for the actual driving behavior (Tang et al. 2018). For highway-rail signalized intersections, a dynamic dilemma zone (i.e., dynamic means the dilemma zone changes with the

vehicle speeds and the number of vehicles on the segment) was proposed based on analyzing a car-following model. Nevertheless, the actual car-following behavior was not considered (Moon and Coleman Fref 2003).

Recent research efforts have revealed that the surrounding infrastructure (i.e., intersections) has a significant impact on the driver car-following behavior when compared to segments in terms of several behavioral measures including following distance, relative speed, headway, and acceleration (Tawfeek and El-Basyouny 2018a, 2019c). Also, Tawfeek and El-Basyouny (Tawfeek and El-Basyouny 2020) used multi-layer neural networks to distinguish driver's car-following intersection-related behavior from segment-related behavior. This neural network was part of a new layer added to driver assistant systems which recognized the variation in car-following behavior due to the change in the surrounding infrastructure (Tawfeek and El-Basyouny 2020).

Therefore, this paper aims at developing a human-like speed model that predicts drivers' speed at the minimum following distance while approaching intersections. This model can be used to enhance the adaptability of AV, from a driver comfort perspective, during a forward collision avoidance situation. To reach this goal, the following objectives should be achieved:

- a) Model the speed at the minimum following distance in car-following during braking, which is considered the most frequent response to a car-following critical safety situation (Adams 1994; Lee et al. 2007),
- b) Investigate the impact of intersection type (i.e., signalized or unsignalized intersection) on the estimated speed, and
- c) Compare the performance of different machine learning techniques to select the most suitable technique for modeling the speed.

In this paper, a total of 12,304 car-following events were extracted from a naturalistic driving study conducted in Ann Arbor, Michigan, US. Six behavioral measures were obtained for each event at different time steps including following distance, relative speed, headway, Time-to-Collision (TTC), following vehicle's acceleration, and speed. Also, the speed at the minimum following distance and the location of the event (i.e., at signalized or unsignalized intersection) were identified. Three different modeling techniques were then adopted to model the speed at the minimum following distance. The developed models were assessed and compared to select a suitable model that can estimate a safe and comfortable speed at the minimum following distance in the vicinity of intersections for AV.

RELATED WORK

Car-following behavior is an extensive area of research due to the high frequency of this particular behavior. Previous studies have presented insights into this behavior that were beneficial for various purposes, such as car-following models' development and calibration (Treiber et al. 2000; Rakha 2009; Tang et al. 2012; Hammit et al. 2018), driver assistance system development (Kiefer et al. 1999; Bella and Russo 2011; Mohammed et al. 2016; Tawfeek and El-Basyouny 2018b), and rear-end near-crash and crash investigation (Carney et al. 2016). Moreover, car-following modeling for autonomous vehicles has been discussed in various studies. For example, Fu et al. (Fu et al. 2019) studied the effect of cut-in vehicles on car-following behavior and drivers' expectations. The cut-in position, gap, reaction time, speed, and acceleration were collected using an instrumented vehicle on a specific freeway during the peak hour. A car-following model for the vehicle following the cut-in vehicle was proposed to enhance the safety and acceptance of autonomous and connected vehicles in a mixed traffic environment.

Moreover, Zhu et al. (Zhu et al. 2018) investigated historical naturalistic driving data to develop a car-following model using deep reinforcement learning for autonomous vehicles. This car-following model mimics human driving in terms of speed, relative speed, following distance, and following vehicle acceleration. Yang et al. (Yang et al. 2019) proposed a combination car-following model that combined machine learning-based and kinematics-based car-following models. The combination method optimized the weight value of these two models. The main goal of this combination was to avoid shortcomings in the machine learning-based car-following models including the replication of potential safety issues that might be learned from the human behavior by the machine learning-based car-following models (Yang et al. 2019).

Chong et al. (Chong et al. 2011) simulated car-following behavior in terms of acceleration using a shallow neural network (i.e., consists of an input layer, one hidden layer, and an output layer); however, the naturalistic data used came from one driver only. The inputs for the model were speed, space headway, and relative speed. Khodayari et al. (Khodayari et al. 2012) also trained a shallow neural network to propose a car-following model, which considered the driver's instantaneous reaction delay. This model predicted the following vehicle acceleration while feeding the relative speed, the following distance, the speed of the following vehicle, and the instantaneous reaction delay to the neural network (Khodayari et al. 2012). As an enhancement to Khodayari et al.'s car-following model, Zheng et al. (Zheng et al. 2013) proposed an instantaneous reaction delay model using neural networks. This reaction delay model was integrated with a car-following model instead of a fixed reaction delay value. The integrated methodology for car-following behavior with the reaction delay outpaced the models that considered fixed reaction delay (Zheng et al. 2013).

He et al. (He et al. 2015) presented a data-driven non-parametric car-following model using k-nearest neighbor. The inputs of the model were the leading and following vehicles' positions in the latest two time steps. These inputs were expressed as the moving distance of the leading vehicle and the space headway of the following vehicle in the latest two time steps, while the output of the model was the following vehicle moving distance. A simulation experiment showed that the proposed model could replicate the real stop-and-go oscillation (He et al. 2015). Also, Dabiri and Abbas (Dabiri and Abbas 2018) developed a data-driven car-following model using gradient boosting regression tree. In general, the proposed model showed superiority to a traditional car-following model (Dabiri and Abbas 2018).

Using a driving simulation, Jian and Shi (Jian and Shi 2020) trained several neural networks to develop car-following models for elderly drivers only, younger drivers only, and both groups. The results showed that the models which were developed for each group separated had slightly better modeling performance. These models were primarily used to quantify the impact of elderly drivers on traffic safety (Jian and Shi 2020). On the other hand, Huang et al. (Huang et al. 2018) investigated the variability in the human driving behavior by investigating the asymmetric driving behavior for hysteresis, discrete driving (i.e., acceleration and deceleration actions are not completely continuous), and intensity difference (i.e., the magnitudes of acceleration and deceleration are different at the same following distance, speed of the following vehicle, and same relative speed magnitude). A car-following was developed using long short-term memory neural networks (Huang et al. 2018).

In summary, there are several studies conducted to model car-following behavior and to cover different aspects of this frequent behavior using a data-driven approach and various machine learning techniques. For example, the instantaneous reaction delay was included and modeled in

the car-following models (Khodayari et al. 2012; Zheng et al. 2013). Moreover, several car-following models were developed to mimic the safe human driving behavior in autonomous vehicles environment (Zhu et al. 2018; Fu et al. 2019; Yang et al. 2019). Also, car-following models were developed for different drivers (i.e., elderly and young drivers) (Jian and Shi 2020) and to model the variability in driver behavior (Huang et al. 2018). Despite these huge efforts, the developed car-following models did not consider the change in the behavior due to the proximity of intersections, especially in the AV environment. Therefore, the main goal of this paper is to model the car-following behavior for AV in terms of following vehicle speed when approaching intersections. Since various machine learning techniques have been used to model the car-following behavior, this paper also investigates the suitability of different machine learning techniques for this task.

METHODOLOGY

Data Description

The naturalistic driving study (NDS) data is considered by many as the most reliable and robust data in terms of closeness to reality when investigating driver behavior. Also, this data covers all network features including all intersection types (e.g., signalized, stop-controlled, etc.). Therefore, this study analyzed a naturalistic driving study from the Safety Pilot Model Deployment (SPMD), which was a data collection effort under real-life conditions in Ann Arbor, Michigan, US (Henclewood and Rajiwade 2015). This dataset is available on the ITS Public Data Hub, which is operated by the Federal Highway Administration (FHWA) under the US DOT (USDOT 2018). The dataset was collected from 64 host vehicles for two months (October 2012 and April 2013)

(Henclewood and Rajiwade 2015). The data includes GPS positions and details, driver behavior data (e.g., speed, acceleration, etc.), and radar data (e.g., following distance, relative speed, etc.).

Despite the robustness and the reliability of the NDS data, it can be challenging to extract target events. Therefore, a driving events extraction process, as discussed in (Tawfeek and El-Basyouny 2018a, 2018b), was followed. In summary, the process started with a data preparation and filtering stage to remove instances with invalid GPS details or any missing information. Also, this stage excluded the vehicles that were identified as trucks, bikes, pedestrians, etc. and only kept the vehicles that were identified as passenger cars. The next stage identified the car-following events during braking in the data, and then the incidents of interests were extracted by scanning the car-following window to select the incidents that occurred at the minimum following distance between the following and the leading vehicles. Finally, ArcGIS was used to map the incidents of interests and to identify the intersection-related events based on a predefined intersection influence area. This area included the physical area of the intersection and the area upstream the stop line of the intersection and up to the stopping sight distance (Tawfeek and El-Basyouny 2018a, 2019c). Moreover, these intersection-related events were classified as signalized and unsignalized intersection events. The unsignalized intersection-related events included stop-controlled or yield-controlled intersection-related events.

Based on the above-mentioned process, a total of 12,304 intersection-related events was identified. These events were split into 2,608 and 9,696 events representing unsignalized and signalized intersections, respectively. The signalized intersections are defined here as any intersection that has 3 or more legs and is controlled with a traffic light. On the other hand, unsignalized intersections are defined as an intersection that is controlled by stop or yield signs. It is worth mentioning that the events occurred on the controlled approaches (i.e., stop-controlled or yield-

controlled) were only included in the analysis. Also, the uncontrolled intersections were excluded from the unsignalized intersection category and any further analysis. For each of these events, several behavioral measures including the following distance, acceleration, speed, relative speed, headway, and TTC were extracted at 1.18 and 1.52 seconds before the following vehicle reaches the minimum following distance. At these times, a warning is desirable to the driver in a car-following situation (i.e., between 1.18 and 1.52 seconds before the minimum following distance) (Kiefer et al. 1999; Bella and Russo 2011). Also, this range is consistent with the time range used to estimate the speed at the yellow onset of the vehicle that leads a platoon approaching signalized intersections (Tawfeek 2020). Therefore, the behavioral measures in this study are extracted at these times to predict the speed before reaching the minimum following distance during braking in the proximity to an intersection. Also, the speed of the following vehicle at the minimum following speed and whether the event was within a signalized or unsignalized intersection influence area were extracted. Table 1 presents summary statistics for these features.

Moreover, the frequency distributions for the speed at the minimum following distance at the signalized intersections and the unsignalized intersections are presented in Figure 1. In order to show the difference in the behavior while driving at these locations, a Kolmogorov-Smirnov Test is conducted, and the results showed that there is a significant difference between the speed at both locations at 99% confidence level. It is also worth noting that the mean and the standard deviation (SD) of the speed at the signalized intersections are 7.03 m/s and 3.59 m/s, respectively. On the other hand, the mean and the SD at the unsignalized intersections are 6.39 m/s and 2.49 m/s, respectively.

Data Reduction

The Principal Component Analysis (PCA) was utilized in this study to extract the maximum amount of information from the features, to reduce the number of features, and to avoid any multicollinearity among these features, which is evident in the correlation matrix in Table 2. The PCA is a commonly used statistical technique that adopts an orthogonal transformation to obtain uncorrelated features from correlated features (Jolliffe 2002). The uncorrelated features, which are the output of the PCA, are called Principal Components (PCs). The first PC includes the highest variability in the data followed by the second PC, and so on. For more details about the PCA, readers are referred to (Jolliffe 2002). The PCs used in this study were selected based on an eigenvalue of equal or greater than 1.

Before performing the PCA, the data was normalized to a zero mean and unit variance. The mean normalization was carried out based on the following equation.

$$\bar{X} = \frac{X - \mu}{\sigma} \quad (1)$$

where $\bar{X}$ is the normalized feature, and μ and σ are the mean and the standard deviation respectively of each feature X .

The PCA was conducted using Python Sickit-learn package (Pedregosa et al. 2011) and the results are summarized in Table 3. As shown in the table, each of the first five PCs had an eigenvalue equal or more than 1, and the five PCs combined account for about 91.42 % of the variance in the data. By investigating the loadings of the five PCs in Table 3, the effect of the behavioral measures on each PC can be interpreted. For example, the two largest loading values in the first PC is for the following distance at 1.52 and 1.18 seconds before the minimum following distance occurs. Hence, it is expected that the following distance has a relatively high impact on the first PC.

Similarly, the highest loading values of the second PC are for the speed followed by the acceleration at 1.52 and 1.18 seconds before the minimum following distance. On the other hand, the third PC is affected by the acceleration at 1.18 seconds and affected inversely (i.e., with a negative sign) by the speed at 1.52 seconds. The fourth PC has a relatively higher weight for the TTC at 1.18 and 1.52 seconds. Finally, the fifth PC has a significantly high loading for the location of the minimum following instance (i.e., at signalized or unsignalized intersection). This brief interpretation of the PCA results will help in understanding the models which will be discussed later in this paper.

Speed Estimation Model Development

Three different machine learning techniques have been adopted in this study to find the best model that can predict the vehicle speed at the minimum following distance in car-following during braking situations in the vicinity of intersections. Several models were developed using linear regression, neural networks (NN), and eXtreme Gradient Boosting (XGBoost) with the aid of Python Sickit-learn package (Pedregosa et al. 2011). For model training and testing purposes, the data was randomly split into a training set (representing 80% of the entire dataset) and a testing set (representing 20% of the entire dataset). The input features for all of the three techniques were the five PCs. The comparison between the best models of each technique was based on the *R-squared* and the Mean Squared Error (MSE) of the test set as Goodness of Fit (GOF) measures.

Linear Regression Model

Linear regression is a common easy-to-interpret modeling technique that has been used in all fields of science and engineering. Linear regression has been used to model driver behavior in similar situations (i.e. car following during braking) in previous studies. For example, Bella & Russo and

Tawfeek & El-Basyouny have used linear regression to model the minimum following distance to develop a Forward Collision Warning (FCW) algorithm using several behavioral measures as dependent variables (e.g. relative speed and acceleration) (Bella and Russo 2011; Tawfeek and El-Basyouny 2018b). Moreover, Tawfeek (2020) has used linear regression to model speed at the yellow onset while approaching signalized intersections and compared the linear regression results to other techniques including neural network (Tawfeek 2020).

In some cases, linear regression shows comparable performance to other techniques including tree-based ensembles and Neural Networks. For example, Wang et al. (2021) adopted several machine learning techniques including Neural Network (multilayer perceptron (MLP) with backpropagation used in training and Relu as an activation function) to predict the aircraft taxi time in three airports (Wang et al. 2021). In their study, the linear regression models outperformed the MLP for all the airports. Furthermore, the performance of linear regression model was substantially close to the performance of the best selected model (e.g., for one of the airports, R -squared equaled 0.82 for the linear regression model and 0.84 for Random Forest model).

Therefore, linear regression was used in this study to predict the speed of the driver at the minimum following distance. The training process of linear regression models is simple and does not need any hyperparameters tuning. The linear regression model performance is evaluated based on the *R-squared* and MSE values.

XGBoost Model

XGBoost (H. Friedman 2001; Chen and Guestrin 2016) is a scalable tree boosting system that has been known recently with its powerful classifications and predictions. XGBoost regression has been adopted in different transportation applications such as taxi demand prediction (Yao et al. 2018), electric vehicles departure time prediction (Frendo et al. 2020), and multistep speed

prediction (Zhang et al. 2019). When compared to other tree-based ensemble methods, XGBoost has shown promising results (Mousa et al. 2018; Mohammadi et al. 2019). Also, the relationship between the input and output variables can be interpreted when XGBoost is used to develop the model. Therefore, it has been used to model the speed at the minimum following distance.

The training procedure of an XGBoost model usually needs hyperparameters tuning such as maximum tree depth, number of trees, training data subsample ratio, and learning rate to reach the best performance possible. Hence, a five-fold cross-validation and grid search were used to fine-tune these hyperparameters. The five-fold cross-validation is used to assess the performance of each combination of the hyperparameters. The cross-validation process starts with splitting the training dataset randomly into roughly equal-sized five groups (i.e., folds). Then, four folds are used for training the model and the fifth is used for testing (or validating) the results. This training and validation process is repeated five times to ensure that each fold has been used once for validation while the other four are used for training (Hastie et al. 2009). The main function of the grid search is simply to develop a model for each hyperparameter combination and to find the best combination, which is defined based on the model performance. The following are the values that were investigated for the hyperparameters fine-tuning process:

- The maximum tree depth ranges from 4 to 20 with an increment of 2,
- The number of trees ranges from 100 to 900 with an increment of 200,
- The training data subsample ratio ranges from 0.7 to 1.0 with an increment of 0.1,
and
- The learning rate values are 0.001, 0.005, 0.01, 0.05, 0.1, 0.25, and 0.35.

The input features for the training process were the five PCs that were generated from the PCA. The model performance was evaluated based on the *R-squared* values for each combination of these hyperparameters and the best combination was selected.

Neural Network Model

Artificial neural networks have been widely used as a modeling and pattern recognition technique in various transportation applications. For instance, NNs were used to investigate the route choice behavior (Yang et al. 1993), model car-following behavior (Jia et al. 2003), driver assistant systems development (Tawfeek and El-Basyouny 2020), traffic congestion prediction (Dougherty et al. 1993), and traffic safety analysis (Tawfeek and El-Basyouny 2019a). In these applications, NN has shown high ability to extrapolate from different data types and make predictions with high accuracies. Therefore, NN is considered as a potential candidate to model the speed at the minimum following distance.

In this study, two-hidden-layer backpropagation feed-forward NNs were trained with a different number of hidden units for each hidden layer (i.e., from 10 to 30 hidden units with an increment of 5 hidden units) to find the best NN. For all the trained networks, the input layer included the five PCs as input features and the output variable was the speed at the minimum following distance. These NNs were trained using Keras libraries in Python and five-fold cross-validation was adopted for each network. The best performing NN was defined as the network that has the minimum MSE of the validation dataset during the cross-validation. However, to compare the NN model to other models, the *R-squared* value was estimated for the best model. The *Relu* activation function was used for the two hidden layers and the Adam optimization, which is a stochastic gradient descent method, was used as the network optimizer.

RESULTS

To model the speed at the minimum following distance in car-following situations during braking, three different techniques (i.e., linear regression, NN, and XGBoost) were utilized to train and select the best model. For all the techniques, the data was split into training and testing datasets to evaluate the performance of the models for each technique. Since there were thirteen features extracted from the data, the PCA was used, as a data preparation step, to reduce the number of features and eliminate any potential multicollinearity in the data. To avoid selecting a model that overfits the data, the comparison between the three techniques are based on the Goodness of Fit measures (including *R-squared* and MSE) of the testing dataset. Table 4 summarizes the performance of the best model for each modeling technique.

Linear Regression Model

Table 5 shows the linear model that was produced using the five PCs. The table also includes the *p-value* for each variable (i.e., the selected five PCs) and it indicates that all the PCs are significant at a 99% confidence level. As shown in the table and based on the coefficients of the PCs, the PC2 has the highest coefficient, which means that it will have the highest contribution to the speed at the minimum following distance. The second highest coefficient magnitude is for PC3, followed by PC4. However, both coefficients for PC3 and PC4 have negative signs, which implies that they are inversely contributing to the speed. Based on the interpretation of the PCA results discussed earlier, speed and acceleration had higher loadings in PC2 when compared to other features. Hence, the speed and acceleration at 1.52 and 1.18 seconds before the minimum following distance have a noticeable effect on the speed at the minimum following distance. Also, the PC5, which depends mainly on the type of intersection (i.e., signalized or unsignalized), has the smallest

coefficient in the model. Nonetheless, the results generally indicate that the intersection type contributes to the speed at the minimum following distance.

The developed linear model has a high performance in terms of the GOF measures as shown in Table 4. The *R-squared* values of the model are 0.953 for the training dataset and 0.959 for the testing dataset, while the MSE values are 0.534 and 0.478 for the training and testing datasets, respectively.

XGBoost Model

A total of 1,120 hyperparameters combinations with a five-fold cross-validation process were used to find the best-trained model. The combination that produced the best *R-squared* value (i.e., the largest value) has 900 trees, a maximum tree depth of 8, a learning rate of 0.01, and a subsample ratio of 0.70. Therefore, the GOF measures of this model were used for comparison with other techniques. As shown in Table 4, the *R-squared* values for the training and testing datasets are 0.99 and 0.971, respectively. Also, the MSE values were calculated for the best model and the results are 0.113 and 0.337 for the training and testing datasets, respectively.

To understand to what extent each PC affects the results of the XGBoost model, the *gain* was estimated for the best model. The gain values reflect how relatively each feature contributes to the model. When the gain value of a feature is higher than the other feature, it is inferred that this feature is relatively important to the model. Figure 2 summarizes the gain percentages of each of the PCs used to model the best XGBoost model. As shown in the figure, the PC2 has the highest gain percentage followed by the PC3, when compared to other PCs, which aligns with the results of the linear regression model.

Neural Network Model

A total of 25 neural networks were explored to find the best NN structure, which produces the best possible accuracy in terms of MSE, which was 0.328 and 0.291 for the training and testing datasets, respectively. The best model has 30 hidden units in the first hidden layer and 25 hidden units in the second layer. The *R-squared* values of the training and testing datasets were estimated as 0.971 and 0.975, respectively as shown in Table 4. Unlike the linear regression and the XGBoost, the NN does not provide a direct indication of the input features' importance.

Models Comparison

Based on the models' training and testing results, as shown in Table 4, the differences between the performances of the models are not significant. However, the performance (in terms of *R-squared* and MSE) of both NN and XGBoost are noticeably better than the linear regression model. When comparing the results of the NN to the XGBoost, it is noted that the XGBoost training performance (i.e., higher *R-squared* and lower MSE) is better than the NN. For the testing dataset, the *R-squared* value of both NN and XGBoost are also close but the MSE value of the NN is slightly less than the XGBoost. This indicates that the NN is marginally better than the XGBoost. Nevertheless, the training time of the NN model is significantly longer (about 12 times more) than the training time of the XGBoost model on the same machine. Therefore, given the marginal difference between the GOF measures for the NN and XGBoost, the XGBoost is more suitable for predicting the speed at the minimum following distance while approaching intersections.

Sensitivity Analysis

The results of the PCA and model development showed that PC2 has the highest gain (in case of XGBoost model) and the highest coefficient magnitude (in the linear regression model), which

means PC2 has the highest impact on the speed at the minimum following distance. According to the PCA results shown in Table 3, it is implied from the loadings of PC2 that the speed at 1.52 and 1.18 seconds before the minimum following distance has the highest impact on PC2. Hence, it is concluded earlier in this study that these speeds have the highest impact on the speed at the minimum following distance. Also, the interpretation of the PCA and model development showed that the type of intersection will impact the speed at the minimum following distance in the vicinity of intersections.

Nevertheless, to better understand the relation between the input and output variables (i.e., the speed at the minimum following distance in the vicinity of intersections) and to confirm the interpretation discussed before, sensitivity analysis has been conducted. The sensitivity analysis is carried out for the selected XGBoost model, which is concluded to be the most suitable model for modeling the speed.

The Sobol variance-based method, which has been selected to conduct the sensitivity analysis in this study, is a recent and effective global sensitivity analysis technique (Sobol I M 2001; Saltelli 2002; Saltelli et al. 2010). This method has shown higher suitability when applied to complex models including transportation-related models (Quaglietta and Punzo 2013; Lima Azevedo et al. 2015). The Sobol variance-based method uses the Monte-Carlo simulation framework to estimate sensitivity indices, and hence, determine the dependencies of an output on various input variables. For more details about the Sobol method, readers are directed to (Sobol I M 2001; Saltelli 2002; Saltelli et al. 2010).

The sensitivity analysis has been conducted using the SALib package in Python and the parameters of the analysis are set as follow:

- The model to be evaluated is coded to include the PCA transformation and then predicting the output based on the selected XGBoost model. In other words, the PCA and the XGBoost are used in the sensitivity analysis after being already trained as discussed before.
- The number of input variables is set to be 13 variables,
- The ranges, which are used as boundaries for the analysis, are set to be equal to the ranges of the variables presented in Table 1, and
- The sample number to be generated is set to be 1000, which generates 28,000 rows to evaluate the XGBoost model.

The results of the analysis are exhibited in Table 6, which summarizes the first-order sensitivity indices for the thirteen input variables and their rank. As shown in the table, the speed at 1.52 and 1.18 seconds before the minimum following distance has the highest impact on the speed at the minimum following distance in the vicinity of intersections. Moreover, the speed at the minimum following distance is sensitive to the headway at 1.52 and 1.18 seconds and the acceleration at 1.18 seconds before the minimum following distance. Finally, the type of intersection is ranked sixth in the variables that influence the speed. On the other hand, the negative indices in Table 6, as in the case of the relative speed and TTC at 1.52 and 1.18 seconds, have no direct impact on the speed. Generally, this result is consistent with the interpretation of PCA and model development results discussed earlier.

DISCUSSION

Three machine learning techniques (i.e., linear regression, NN, and XGBoost) were used to model the driver speed at the minimum following distance while approaching intersections. The XGBoost was concluded to be the most suitable machine learning technique to predict driver speed. This model can be applied in various systems with different automation levels to ensure driver's riding

comfort while meeting the surrounding human drivers expectations in fully autonomous vehicles or to provide the driver with a reliable advisory safe speed in semi-autonomous vehicles.

The proposed model can be used with various Connected Autonomous Vehicles applications such as Forward Collision Warning (FCW) or other V2I applications. For the FCW, research has shown that the car-following behavior (including the following distance between the leading and following vehicles) on segment changes significantly when compared to the driver behavior in the vicinity of intersections (Tawfeek and El-Basyouny 2019c, 2020). Accordingly, and since most of the FCW algorithms depend on the following distance to warn the driver about an impending rear-end collision (Tawfeek and El-Basyouny 2018b), it is expected that the same driver will perceive this warning differently when driving in the vicinity of intersections. Therefore, the proposed model can be used when a rear-end incident is detected to advise the driver to drive at a safe speed (in a semi-autonomous vehicle environment) or recommend a comfortable speed to the system (in a fully AV environment). This safe or comfortable speed should be the maximum speed for the driver or the system in such a situation.

For V2I application, various warnings or advisory messages can be provided to the driver based on the information shared from a road side unit which is typically connected to the traffic signal controller at signalized intersections. Some of these warnings and messages aim at minimizing fuel consumption and emissions (Kamalanathsharma and Rakha 2013) or helping the driver to safely navigate the dilemma zone (Xie and Wang 2018; Tawfeek 2020). However, the algorithms that decide on these warnings or advisory messages did not explicitly investigate the behavior of the following vehicle, and hence, these algorithms did not provide clear guidance or advice to the following vehicle. Therefore, the model proposed in the current study extends the efforts conducted in the literature by guiding the following vehicle. This guidance is defined here as a

safe speed that can help the driver avoid rear-end conflict or evasive maneuver. For example, this safe speed can be provided to the following vehicle when the leading vehicle decides to decelerate, based on a V2I communication, to avoid stopping at the red light. In a fully AV environment, the proposed model can recommend a comfortable speed, from a driver perspective, when the leading vehicle needs to slow down.

SUMMARY AND CONCLUSION

This study adopted three different machine learning techniques (i.e., linear regression, NN, and XGBoost) to model the speed at the minimum following distance while approaching intersections. The model development process started with applying PCA on thirteen selected variables to eliminate any potential multicollinearity in the data. These variables included the following distance, relative speed, headway, and TTC between the following and leading vehicles, and the speed and acceleration of the following vehicle. These six measures were extracted from the NDS data at 1.52 and 1.18 seconds before reaching the minimum following distance. Also, the intersection control type (i.e., signalized or unsignalized) where the minimum following distance took place was identified using ArcGIS. A total of 12,304 car-following events were used for the model development process.

Based on the PCA, the number of features input to the models was reduced from thirteen features to five uncorrelated PCs, which are then used as input features in the model development process. A five-fold cross-validation process with a grid search was adopted to fine-tune the hyperparameters for XGBoost models and to find the best model structure for the NN models. The criteria for selecting the best model were based on the GOF measures (i.e., *R-squared* and MSE) of the testing dataset. The results indicate that the GOF measures of the testing dataset for the XGBoost and NN outpace the linear regression. Although there is a marginal difference between

the XGBoost and NN GOF measures, the training time of the NN models was about 12 times more than the training time of the XGBoost. Therefore, the XGBoost model is identified as the most suitable technique to model the speed at the minimum following distance. Also, the importance of the input features was investigated for both XGBoost and linear regression models. The result exhibits that the intersection control type (i.e., signalized or unsignalized) contributes to the speed at the minimum following distance. This observation highlights that during the development of human-like car-following models for AV, it is crucial to consider not only the impact of intersections on the car-following behavior but also the type of intersection control.

This paper has examined different machine learning techniques to predict the speed at the minimum following distance for AV human-like control. Future research should extend the assessment of the best model by comparing it to other car-following models. Also, it is important to evaluate the impact of the proposed model on safety, mobility, and environment in a simulation environment with varied market penetration rates as suggested in (Tawfeek and El-Basyouny 2019b). Moreover, the proposed model can be used to enhance the calibration process of microsimulation models.

REFERENCES

- Adams, L. 1994. Review of the literature on obstacle avoidance maneuvers: Braking versus steering. (August): 22. doi:10.1186/1757-1626-2-9342.7.
- Bella, F., and Russo, R. 2011. A Collision Warning System for rear-end collision : a driving simulator study. *Procedia - Social and Behavioral Sciences*, **20**: 676–686. doi:10.1016/j.sbspro.2011.08.075.
- Bevrani, K., and Chung, E. 2012. A safety adapted car following model for traffic safety studies.

- Advances in Human Aspects of Road and Rail Transportation, 550–559. doi:10.1201/b12320.
- Carney, C., Harland, K.K., and McGehee, D. V. 2016. Using event-triggered naturalistic data to examine the prevalence of teen driver distractions in rear-end crashes. *Journal of Safety Research*, **57**: 47–52. doi:10.1016/j.jsr.2016.03.010.
- Chen, T., and Guestrin, C. 2016. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, **13-17-Aug**: 785–794. doi:10.1145/2939672.2939785.
- Chong, L., Abbas, M.M., and Medina, A. 2011. Simulation of driver behavior with agent-based back-propagation neural network. *Transportation Research Record*, (2249): 44–51. doi:10.3141/2249-07.
- Dabiri, S., and Abbas, M. 2018. Evaluation of the Gradient Boosting of Regression Trees Method on Estimating Car-Following Behavior. *Transportation Research Record*, **2672**(45): 136–146. doi:10.1177/0361198118772689.
- Dougherty, M.S., Kirby, H.R., and Boyle, R.D. 1993. The use of neural networks to recognise and predict traffic congestion. *Traffic engineering & control*, **34**(6): 311–4.
- Frendo, O., Gaertner, N., and Stuckenschmidt, H. 2020. Improving Smart Charging Prioritization by Predicting Electric Vehicle Departure Time. *IEEE Transactions on Intelligent Transportation Systems*, 1–8. doi:10.1109/tits.2020.2988648.
- Fu, R., Li, Z., Sun, Q., and Wang, C. 2019. Human-like car-following model for autonomous vehicles considering the cut-in behavior of other vehicles in mixed traffic. *Accident Analysis and Prevention*, **132**. doi:10.1016/j.aap.2019.105260.

- H. Friedman, J. 2001. Greedy Function Approximation: a Gradient Boosting Machine. *Annals of Statistics*, (2): 34.
- Hammit, B.E., Ghasemzadeh, A., James, R.M., Ahmed, M.M., and Young, R.K. 2018. Evaluation of weather-related freeway car-following behavior using the SHRP2 naturalistic driving study database. *Transportation Research Part F: Traffic Psychology and Behaviour*, **59**: 244–259. doi:10.1016/j.trf.2018.08.023.
- Hastie, T., Tibshirani, R., and Friedman, J. 2009. The Elements of Statistical Learning. *In* Springer series in statistics, Second Edi.
- He, Z., Zheng, L., and Guan, W. 2015. A simple nonparametric car-following model driven by field data. *Transportation Research Part B: Methodological*, **80**: 185–201. doi:10.1016/j.trb.2015.07.010.
- Henclewood, D., and Rajiwade, S. 2015. Safety Pilot Model Deployment – Sample Data Environment Data Handbook. **1.3**.
- Huang, X., Sun, J., and Sun, J. 2018. A car-following model considering asymmetric driving behavior based on long short-term memory neural networks. *Transportation Research Part C: Emerging Technologies*, **95**: 346–362. doi:10.1016/j.trc.2018.07.022.
- Jia, H., Juan, Z., and Ni, A. 2003. Develop a car-following model using data collected by “five-wheel system.” *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, **1**: 346–351. doi:10.1109/ITSC.2003.1251975.
- Jian, M., and Shi, J. 2020. Analysis of impact of elderly drivers on traffic safety using ANN based car-following model. *Safety Science*, **122**. doi:10.1016/j.ssci.2019.104536.
- Jolliffe, I.T. 2002. *Principal Component Analysis*. Springer-Verlag, New York.

- Kamalanathsharma, R.K., and Rakha, H.A. 2013. Multi-stage dynamic programming algorithm for eco-speed control at traffic signalized intersections. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC.; 2094–2099. doi:10.1109/ITSC.2013.6728538.
- Khodayari, A., Ghaffari, A., Kazemi, R., and Brauningstingl, R. 2012. A modified car-following model based on a neural network model of the human driver effects. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, **42**(6): 1440–1449. doi:10.1109/TSMCA.2012.2192262.
- Kiefer, R., LeBlanc, D., Palmer, M., Salinger, J., Deering, R., and Shulman, M. 1999. Development and Validation of Functional Definitions and Evaluation Procedures For Collision Warning/Avoidance Systems. (August): 75. doi:Report No. DOT-HS-808-964.
- Lee, S.E., Llaneras, E., Klauer, S.G., and Sudweeks, J. 2007. Analyses of Rear-End Crashes and Near-Crashes in the 100-Car Naturalistic Driving Study to Support Rear-Signaling Countermeasure Development. Distribution, (October): 1–125.
- Lima Azevedo, C., Ciuffo, B., Cardoso, J.L., and Ben-Akiva, M.E. 2015. Dealing with uncertainty in detailed calibration of traffic simulation models for safety assessment. Transportation Research Part C: Emerging Technologies, **58**: 395–412. doi:10.1016/j.trc.2015.01.029.
- Mohammadi, R., He, Q., Ghofrani, F., Pathak, A., and Aref, A. 2019. Exploring the impact of foot-by-foot track geometry on the occurrence of rail defects. Transportation Research Part C: Emerging Technologies, **102**: 153–172. doi:10.1016/j.trc.2019.03.004.
- Mohammed, M., Ke, Y., Gao, J., Zhang, H., El-Basyouny, K., and Qiu, T.Z. 2016. Connected vehicle V2I communication application to enhance driver awareness at signalized intersections. Proceedings, Annual Conference - Canadian Society for Civil Engineering,

London, Ontario, Canada, Vol. 4.

- Moon, Y.J., and Coleman Fref, I. 2003. Dynamic dilemma zone based on driver behavior and car-following model at highway-rail intersections. *Transportation Research Part B: Methodological*, **37**(4): 323–344. doi:10.1016/S0191-2615(02)00014-0.
- Mousa, S.R., Bakhit, P.R., Osman, O.A., and Ishak, S. 2018. A comparative analysis of tree-based ensemble methods for detecting imminent lane change maneuvers in connected vehicle environments. *Transportation Research Record*, **2672**(42): 268–279. doi:10.1177/0361198118780204.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**: 2825–2830.
- Przybyla, J., Taylor, J., Jupe, J., and Zhou, X. 2012. Simplified, data-driven, errorable car-following model to predict the safety effects of distracted driving. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*,: 1149–1154. doi:10.1109/ITSC.2012.6338913.
- Quaglietta, E., and Punzo, V. 2013. Supporting the design of railway systems by means of a Sobol variance-based sensitivity analysis. *Transportation Research Part C: Emerging Technologies*, **34**: 38–54. doi:10.1016/j.trc.2013.05.007.
- Rakha, H. 2009. Validation of Van Aerde's simplified steadystate car-following and traffic stream model. *Transportation Letters*, **1**(3): 227–244. doi:10.3328/TL.2009.01.03.227-244.
- Saifuzzaman, M., and Zheng, Z. 2014. Incorporating human-factors in car-following models : A

- review of recent developments and research needs. *Transportation Research Part C*, **48**: 379–403. doi:10.1016/j.trc.2014.09.008.
- Saltelli, A. 2002. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, **145**(2): 280–297. doi:10.1016/S0010-4655(02)00280-1.
- Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S. 2010. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, **181**(2): 259–270. doi:10.1016/j.cpc.2009.09.018.
- Sobol I M. 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and computers in simulation*, **55**(1): 271–280.
- Tang, T., Wang, Y., Yang, X., and Wu, Y. 2012. A new car-following model accounting for varying road condition. *Nonlinear Dynamics*, **70**(2): 1397–1405. doi:10.1007/s11071-012-0542-8.
- Tang, T.Q., Yi, Z.Y., Zhang, J., Wang, T., and Leng, J.Q. 2018. A speed guidance strategy for multiple signalized intersections based on car-following model. *Physica A: Statistical Mechanics and its Applications*, **496**: 399–409. doi:10.1016/j.physa.2018.01.005.
- Tawfeek, M.H. 2020. Perceptual-based driver behaviour modelling at the yellow onset of signalised intersections. *Journal of Transportation Safety & Security*,: 1–26. doi:10.1080/19439962.2020.1783414.
- Tawfeek, M.H., and El-Basyouny, K. 2018a. Using Naturalistic Driving Data to Quantify Driver Following Behavior During Braking. *Proceedings, Annual Conference - Canadian Society for Civil Engineering*, Fredericton, NB, Canada.
- Tawfeek, M.H., and El-Basyouny, K. 2018b. A perceptual forward collision warning model using

- naturalistic driving data. *Canadian Journal of Civil Engineering*, **45**(10): 899–907.
doi:10.1139/cjce-2017-0592.
- Tawfeek, M.H., and El-Basyouny, K. 2019a. Estimating Traffic Volume on Minor Roads at Rural Stop-Controlled Intersections using Deep Learning. *Transportation Research Record: Journal of the Transportation Research Board*,. doi:10.1177/0361198119837236.
- Tawfeek, M.H., and El-Basyouny, K. 2019b. Network-level comparison of various Forward Collision Warning algorithms. *SIMULATION*, **95**(4): 313–325.
doi:10.1177/0037549718777613.
- Tawfeek, M.H., and El-Basyouny, K. 2019c. Location-based analysis of car-following behavior during braking using naturalistic driving data. *Canadian Journal of Civil Engineering*,: 1–8.
doi:10.1139/cjce-2019-0314.
- Tawfeek, M.H., and El-Basyouny, K. 2020. A context identification layer to the reasoning subsystem of context-aware driver assistance systems based on proximity to intersections. *Transportation Research Part C: Emerging Technologies*, **117**: 102703.
doi:10.1016/j.trc.2020.102703.
- Treiber, M., Hennecke, A., and Helbing, D. 2000. Congested traffic states in empirical observations and microscopic simulations. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, **62**(2): 1805–1824.
doi:10.1103/PhysRevE.62.1805.
- USDOT. 2018. ITS Public Data Hub. Available from <https://www.its.dot.gov/data/>. [accessed 3 January 2018].
- Wang, X., Brownlee, A.E.I., Woodward, J.R., Weiszer, M., Mahfouf, M., and Chen, J. 2021.

- Aircraft taxi time prediction: Feature importance and their implications. *Transportation Research Part C: Emerging Technologies*, **124**. doi:10.1016/j.trc.2020.102892.
- Wang, X., Zhu, M., Chen, M., and Tremont, P. 2016. Drivers' rear end collision avoidance behaviors under different levels of situational urgency. *Transportation Research Part C: Emerging Technologies*, **71**: 419–433. doi:10.1016/j.trc.2016.08.014.
- Xie, X.F., and Wang, Z.J. 2018. SIV-DSS: Smart In-Vehicle Decision Support System for driving at signalized intersections with V2I communication. *Transportation Research Part C: Emerging Technologies*, **90**: 181–197. doi:10.1016/j.trc.2018.03.008.
- Yang, D., Zhu, L., Liu, Y., Wu, D., and Ran, B. 2019. A Novel Car-Following Control Model Combining Machine Learning and Kinematics Models for Automated Vehicles. *IEEE Transactions on Intelligent Transportation Systems*, **20**(6): 1991–2000. doi:10.1109/TITS.2018.2854827.
- Yang, H., Kitamura, R., Jovanis, P.R., Vaughn, K.M., and Abdel-Aty, M.A. 1993. Exploration of route choice behavior with advanced traveler information using neural network concepts. *Transportation*, **20**: 199–223. doi:10.1007/BF01307059.
- Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., Gong, P., Li, Z., Ye, J., and Chuxing, D. 2018. Deep multi-view spatial-temporal network for taxi demand prediction. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018,: 2588–2595.
- Yu, S., and Shi, Z. 2015. Analysis of car-following behaviors considering the green signal countdown device. *Nonlinear Dynamics*, **82**(1–2): 731–740. doi:10.1007/s11071-015-2191-1.
- Zhang, Z., Li, M., Lin, X., Wang, Y., and He, F. 2019. Multistep speed prediction on traffic

networks: A deep learning approach considering spatio-temporal dependencies. *Transportation Research Part C: Emerging Technologies*, **105**: 297–322. doi:10.1016/j.trc.2019.05.039.

Zheng, J., Suzuki, K., and Fujita, M. 2013. Car-following behavior with instantaneous driver-vehicle reaction delay: A neural-network-based methodology. *Transportation Research Part C: Emerging Technologies*, **36**: 339–351. doi:10.1016/j.trc.2013.09.010.

Zhu, M., Wang, X., and Wang, Y. 2018. Human-like autonomous car-following model with deep reinforcement learning. *Transportation Research Part C: Emerging Technologies*, **97**: 348–368. doi:10.1016/j.trc.2018.10.024.

Draft

LIST OF FIGURES

FIGURE 1 The frequency distribution of the speed at the minimum following distance for both intersection types (i.e., signalized and unsignalized intersections).

FIGURE 2 The significance of each PC based on the gain.

Draft

Table 1 Summary Statistics of the Extracted Behavioral Measures

	Mean	SD	Minimum	Maximum
At 1.52 seconds before reaching the minimum following distance				
Following Distance (m)	13.76	6.87	1.60	63.87
Acceleration (m/s^2)	-1.49	0.86	-5.05	2.63
Speed (m/s)	8.68	3.18	4.52	26.20
Relative Speed (m/s)	-3.43	2.18	-18.50	1.10
Headway (seconds)	1.63	0.59	0.15	6.35
TTC (seconds)	6.77	11.95	0.71	470.08
At 1.18 seconds before reaching the minimum following distance				
Following Distance (m)	14.81	7.11	1.66	66.50
Acceleration (m/s^2)	-1.42	0.88	-4.64	2.03
Speed (m/s)	9.08	3.15	4.48	33.24
Relative Speed (m/s)	-3.67	2.35	-18.80	2.30
Headway (seconds)	1.67	0.60	0.15	7.66
TTC (seconds)	6.99	12.04	0.81	304.64
At the minimum following distance				
Signalized/unsignalized intersection (0/1)	0.21	0.41	0	1
Speed (m/s)	6.89	3.39	4.47	24.17

Table 2 Correlation Matrix between the variables

		At 1.18 Seconds						At 1.52 Seconds						Signalized/ unsignalized	Speed at Minimum Following Distance
		Following Distance	Accel	Speed	Relative Speed	H	TTC	Following Distance	Accel	Speed	Relative Speed	H	TTC		
At 1.18 Seconds	Following Distance	1.00	0.10	0.59	-0.33	0.63	0.21	1.00	0.09	0.59	-0.30	0.63	0.18	-0.07	0.58
	Acceleration (Accel)	0.10	1.00	0.12	0.33	0.03	0.21	0.06	0.92	0.05	0.37	0.05	0.24	-0.04	0.32
	Speed	0.59	0.12	1.00	0.05	-0.18	0.29	0.56	0.15	0.99	0.06	-0.19	0.26	-0.06	0.96
	Relative Speed	-0.33	0.33	0.05	1.00	-0.44	0.37	-0.41	0.33	0.02	0.98	-0.50	0.36	-0.04	0.15
	Headway (H)	0.63	0.03	-0.18	-0.44	1.00	-0.01	0.65	-0.01	-0.18	-0.41	0.99	-0.01	-0.05	-0.15
	TTC	0.21	0.21	0.29	0.37	-0.01	1.00	0.17	0.20	0.27	0.36	-0.04	0.72	-0.05	0.34
At 1.52 Seconds	Following Distance	1.00	0.06	0.56	-0.41	0.65	0.17	1.00	0.06	0.56	-0.38	0.65	0.15	-0.06	0.55
	Acceleration	0.09	0.92	0.15	0.33	-0.01	0.20	0.06	1.00	0.09	0.37	0.01	0.23	-0.04	0.32
	Speed	0.59	0.05	0.99	0.02	-0.18	0.27	0.56	0.09	1.00	0.02	-0.19	0.24	-0.06	0.94
	Relative Speed	-0.30	0.37	0.06	0.98	-0.41	0.36	-0.38	0.37	0.02	1.00	-0.47	0.39	-0.05	0.16
	H	0.63	0.05	-0.19	-0.50	0.99	-0.04	0.65	0.01	-0.19	-0.47	1.00	-0.04	-0.05	-0.15
	TTC	0.18	0.24	0.26	0.36	-0.01	0.72	0.15	0.23	0.24	0.39	-0.04	1.00	-0.05	0.31
Signalized/unsignalized		-0.07	-0.04	-0.06	-0.04	-0.05	-0.05	-0.06	-0.04	-0.06	-0.05	-0.05	-0.05	1.00	-0.08
Speed at Minimum Following Distance		0.58	0.32	0.96	0.15	-0.15	0.34	0.55	0.32	0.94	0.16	-0.15	0.31	-0.08	1.00

Table 3 PCA Results Summary

	Eigenvalue	Variance %	Cumulative variance %		PCs Loadings					
					PC1	PC2	PC3	PC4	PC5	
PC1	4.1	31.3%	31.27%	At 1.18 seconds before the minimum following distance	Following Distance (m)	0.429	0.243	-0.022	-0.003	0.009
PC2	3.5	26.7%	57.94%		Acceleration (m/s^2)	-0.064	0.301	0.424	-0.428	0.029
PC3	2.1	15.8%	73.74%		Speed (m/s)	0.141	0.393	-0.400	-0.184	-0.008
PC4	1.3	10.1%	83.78%		Relative Speed (m/s)	-0.355	0.269	0.113	0.132	-0.043
PC5	1.0	7.6%	91.42%		Headway (seconds)	0.394	-0.062	0.365	0.163	-0.008
PC6	0.7	5.0%	96.43%		TTC (seconds)	-0.027	0.359	0.072	0.493	0.098
PC7	0.3	2.2%	98.61%	At 1.52 seconds before the minimum following distance	Following Distance (m)	0.447	0.210	-0.033	-0.013	0.012
PC8	0.1	0.6%	99.19%		Acceleration (m/s^2)	-0.070	0.305	0.394	-0.455	0.028
PC9	0.1	0.6%	99.77%		Speed (m/s)	0.148	0.372	-0.433	-0.156	-0.008
PC10	0.02	0.2%	99.94%		Relative Speed (m/s)	-0.345	0.281	0.137	0.114	-0.045
PC11	0.01	0.04%	99.98%		Headway (seconds)	0.403	-0.076	0.367	0.119	-0.001
PC12	0.002	0.01%	99.998%		TTC (seconds)	-0.036	0.356	0.101	0.484	0.094
PC13	0.0002	0.002%	100.00%	Signalized/unsignalized intersection		-0.024	-0.062	-0.032	-0.059	0.988

TABLE 4 GOF Comparison between the Three Modelling Techniques

GOF Measure	Dataset	Linear Regression	NN	XGB
MSE	Training	0.534	0.328	0.113
	Testing	0.478	0.291	0.337
R2	Training	0.953	0.971	0.990
	Testing	0.959	0.975	0.971

Draft

TABLE 5 Linear Regression Model Parameters

	Estimate	p-value
Intercept	6.897	<0.001
PC1	0.403	<0.001
PC2	1.490	<0.001
PC3	-0.939	<0.001
PC4	-0.761	<0.001
PC5	-0.035	<0.001

Draft

TABLE 6 Sensitivity Analysis Results

		1st Order Sensitivity Index	Rank
At 1.18 Seconds	Following Distance	0.025	7
	Acceleration	0.049	5
	Speed	0.129	2
	Relative Speed	-0.002	11
	Headway (H)	0.081	4
	TTC	0.008	10
At 1.52 Seconds	Following Distance	0.015	9
	Acceleration	0.016	8
	Speed	0.251	1
	Relative Speed	-0.004	12
	H	0.102	3
	TTC	-0.016	13
Signalized/unsignalized		0.033	6

FIGURE 1 The frequency distribution of the speed at the minimum following distance for both intersection types (i.e., signalized and unsignalized intersections).

FIGURE 2 The significance of each PC based on the gain.