A Simulation Study of Predicting Real-Time Conflict-Prone Traffic Conditions

Christos Katrakazas, *Member, IEEE*, Mohammed Quddus, and Wen-Hua Chen, *Senior Member, IEEE*

***Abstract*— Current approaches to estimate the probability of a traffic collision occurring in real-time primarily depend on comparing traffic conditions just prior to collisions with normal traffic conditions. Most studies acquire pre-collision traffic conditions by matching the collision time in the national crash database with the time in the traffic database. Since the reported collision time sometimes differs from the actual time, the matching method may result in traffic conditions not representative of pre-collision traffic dynamics. In this paper, this is overcome through the use of highly disaggregated vehicle-based traffic data from a traffic micro-simulation (i.e., VISSIM) and the corresponding traffic conflicts data generated by the surrogate safety assessment model (SSAM). In particular, the idea is to use traffic conflicts as surrogate measures of traffic safety so that traffic collisions data are not needed. Three classifiers (i.e., support vector machines, k-nearest neighbours, and random forests) are then employed to examine the proposed idea. Substantial efforts are devoted to making the traffic simulation as representative of the real-world as possible by employing data from a motorway section in England. Four temporally aggregated traffic datasets (i.e., 30 s, 1 min, 3 min, and 5 min) are examined. The main results demonstrate the viability of using traffic micro-simulation along with the SSAM for real-time conflicts prediction and the superiority of random forests with 5-min temporal aggregation in the classification results. However, attention should be given to the calibration and validation of the simulation software so as to acquire more realistic traffic data, resulting in more effective prediction of conflicts.**

***Index Terms*— Traffic safety, traffic conflicts, traffic micro-simulation, support vector machines (SVMs), k-nearest neighbours (k-NN), random forests (RFs).**

I. Introduction

IN RECENT years, the estimation of *unsafe* traffic conditions in real-time has been studied by many Intelligent Transport Systems (ITS) experts. The significance of real-time collision prediction is related to its integral part within a proactive highway safety management that has the potential to reduce road traffic fatalities and injuries. In particular, predicting where and when a traffic collision is likely to occur in real-time and preventing the collision by adjusting the traffic dynamics through a range of traffic management interventions (e.g. variable message signs) are beneficial to highway safety. Previous research on this topic has established an underpinning theory suggesting that traffic dynamics (e.g. interactions between speed, flow and congestion) and spatio-temporal collision risk are highly correlated [1]. Based on this principle, a dominating approach for detecting *unsafe* traffic conditions is the comparison of traffic situations just prior to traffic collision occurrences on a segment with the traffic conditions at normal situations on the same segment. In the current era where various advanced driver assistance systems [2] and autonomous vehicles [3] are massively developed, it becomes essential to effectively identify these traffic fluctuations in real-time and enhance collision-freed decision making of such technologies.

Perhaps the most important factor in the development of real-time collision-prone traffic conditions models is the temporal aggregation of traffic data which would lead to the correct distinction between collision-prone and normal traffic conditions. Temporal aggregation of traffic data is available at pre-defined time intervals (e.g. *30*-second or *1*-minute, *5*-minute and *15*-minute) from the corresponding traffic agencies. Highly disaggregated traffic data (e.g. *30*-second or *1*-minute of temporal aggregation) are usually considered not suitable for implementing a timely intervention by the relevant authorities to intervene and prevent both the collision and the collision-related congestion. This is due to the fact that in the majority of recent studies [4]–[6], traffic conditions at 5-10 minutes before the collision have been found to be the most suitable time period to identify such events timely and initiate an intervention by the responsible traffic agencies. On the other hand, highly disaggregated traffic data may not be available in many countries. Furthermore, even if highly disaggregated traffic data are available, an error exists between the reported collision time and the actual time of a collision. This is because the reported time and location largely depend on the subjective volition of the police officers attending the site of the collision [7]. As a result, inaccurately reported collision time leads to misrepresentative pre-collision traffic dynamics resulting in an inaccurate calibration of the collision prediction models [7].

The inherent difficulties with the recorded collision time and temporal data aggregation issues can be overcome using traffic micro-simulation. Recent research on traffic micro-simulation and road safety (e.g. [8], [9]) showed that it is now possible to estimate surrogate measures of safety performance based on dangerous vehicle interactions. If these

Manuscript received December 2, 2016; revised June 27, 2017; accepted October 24, 2017. Date of publication December 8, 2017; date of current version October 3, 2018. This work was supported by the U.K. Engineering and Physical Sciences Research Council under Grant EP/J011525/1. The Associate Editor for this paper was Y. Wang. *(Corresponding author: Christos Katrakazas.)*

C. Katrakazas and M. Quddus are with the School of Civil and Building Engineering, Loughborough University, Loughborough LE11 3TU, U.K. (e-mail: c.katrakazas@lboro.ac.uk; m.a.quddus@lboro.ac.uk).

W.-H. Chen is with the Aeronautical and Automotive Engineering Department, Loughborough University, Loughborough LE11 3TU, U.K. (e-mail: w.chen@lboro.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2017.2769158

risky vehicle interactions are filtered with established risk indicating thresholds, they are termed as "traffic conflicts". According to the definition by Amundsen and Hyden [10] traffic conflicts occur when two or more road vehicles are in such a collision course that a high probability of a collision exists if their motion remains uninterrupted.

Using traffic conflicts can, therefore, address the issues related to traffic collisions as discussed above. Furthermore, studying conflicts can enhance the understanding of the specific characteristics that lead road users to drive unsafely and cause collisions [8]. Approaches that use traffic conflicts are however criticised in the literature because the correlation between traffic conflicts and traffic collisions on a segment may be low [9]. Nevertheless, it is admitted that the mechanism that triggers collisions and conflicts is analogous [8], [9].

Additionally, because of the technological advances in the area of automated driving, the concept of real-time collision prediction should not necessarily relate to a timely intervention from traffic authorities but rather should concentrate on improving the speed of the prediction. In that way network-level collision prediction can be taken into account for vehicle-level risk assessment. Therefore, the exploitation of highly disaggregated traffic data should be taken into account. In that direction machine learning and data mining approaches can prove advantageous over traditional (e.g. logistic regression [11]) or sophisticated techniques (e.g. Neural Networks [12]) for real-time collision prediction. Since collisions and conflicts are more rare events than normal traffic conditions, attention should also be given to the handling of imbalanced data [13] (i.e. data used for the classification where one class has significantly more instances than the other).

The combination of traffic micro-simulation and machine learning classifiers to detect conflict-prone traffic conditions on motorways from highly disaggregated data form the motivation for the current paper. This study explores the application of commonly employed Support Vector Machines (SVM), *k*-Nearest Neighbours (*k*NN) a simple but effective non-parametric classifier and Random Forests (RF) an ensemble classifier for the classification of simulated traffic data with regards to traffic conflicts. The traffic data used in this study come from the PTV VISSIM micro-simulation software [14] and consist of speed, flow and acceleration data aggregated at different temporal units (e.g. *30*-second, *1*-minute, *3*-minute and *5*-minute time intervals) to compare the effectiveness of the temporal aggregation on the classification results. The conflict data are acquired through the Surrogate Safety Assessment Model (SSAM) [15], a software which uses the trajectories of the vehicles from the traffic micro-simulation and outputs traffic conflicts. A matched-case control data structure is used, in which traffic conditions before each conflict is matched with normal traffic conditions coming from three other simulation runs. The number of three additional runs was chosen in order to cope with the imbalance between conflict and safe conditions which can prove essential for classification purposes [13].

The rest of the paper is organised as follows: firstly, the existing literature and its main findings are synthesised. An analytic description of SVM, kNN and RF classification algorithms is described next. This is followed by a presentation of the data used in the analysis along with the pre-processing methodology and the results of the classification algorithms. Finally, the last section summarises the main conclusions of the study and offers some recommendations for future research.

II. Literature Review

The purpose of this review was to synthesize existing studies on safety assessment using traffic conflicts by comparing and contrasting their findings and identify whether there is any important or interesting knowledge gap. Focus was also given on the methods employed in real-time collision prediction algorithms so as to select the appropriate methods for predicting conflict-prone traffic conditions.

A. Safety Assessment Using Simulated Conflicts

The use of traffic conflicts in road safety assessment using traffic micro-simulation has gained popularity within the ITS research community over the recent years.

In all traffic microsimulation platforms, simulating traffic collisions is not possible because such software is programmed according to a number of safety-related parameters. These parameters include the free-flow speed of cars, inter-vehicle headways, acceleration or deceleration profiles, the interaction between priority and non-priority vehicles, appropriate overtaking and lane-changing gaps as well as the obedience of traffic regulations [16]. Despite these safety related constraints, the fact that vehicles can come very close to each other and the information on vehicles' exact positions, speeds, headings and accelerations can provide a relevant safety index for vehicle interactions [17].

Minderhoud and Bovy [18] suggested that traffic micro-simulation can overcome the need to collect collision data and also provide alternatives to the safety evaluation of ITS technologies. They indicated that safety indexes such as Time-to-Collision (TTC) and the vehicles' headway distribution as provided by traffic microsimulation software can reveal safe and unsafe driving patterns. Likewise, Archer [19] stated that the traffic conflict technique based on the results from micro-simulation could have a practical impact and provide an insight on the identification of safety problems in real-world traffic environments. In order to assess safety within traffic microsimulation environments, Gettman *et al.* [20] investigated the potential of detecting traffic conflicts from surrogate safety indicators such as TTC, Post-Encroachment-Time (PET), the maximum speed of the vehicles, the deceleration rate and the speed differential between the vehicles. Their work was reflected in the development of SSAM, a post-processing software which investigates simulated vehicle trajectories and detects the number and severity of traffic conflicts accompanied by surrogate safety measures for each conflict. Currently SSAM is probably the only exceptional tool for exploiting traffic conflicts from microsimulation [21].

The convenience in terms of the reduced need for on-field data collection and the relatively easy identification of hazardous vehicle encounters through safety indices led to a number of safety-related microsimulation studies. A detailed overview of approaches concerning safety-related traffic simulation was published by Young *et al.* [22]. In their review, it is revealed that researchers are looking to establish a correlation between the number of simulated conflicts with the number of expected real-world collisions. El-Basyouny and Sayed [8] justified the attempt to link conflicts with collisions by indicating that conflicts are based on vehicle interactions compared to typical collision predictors such as exposure. Essa and Sayed [23] and Huang *et al.* [21] however emphasised that the link between conflicts and collisions depends heavily on the calibration of the simulation model. In the same principle, Fan *et al.* [24], who investigated the safety of motorway merging areas, suggested that SSAM should be used with caution because of the purely stochastic nature of real-world collisions.

However, a thorough examination of papers attempting to link conflicts with collisions (e.g. [8], [9], [25]) reveals that the primary aim of these papers is the before-and-after evaluation of new technologies or infrastructure modifications with regards to safety. More specifically, these approaches seek to estimate if alterations to the current state of (a part of) the traffic environment will increase or reduce the number of collisions on specific spots. For instance, a recent study from Shahdah *et al.* [9] used VISSIM with SSAM to develop a statistical relationship between conflicts and collisions for signalised intersections. Traffic conflicts were estimated by using two thresholds for TTC (i.e. 1.5 and 0.5 seconds) and the simulated conflicts were used to calculate crash modification factors (CMFs) for before and after analyses of untreated intersections.

Consequently, an emerging research gap is that of using the simulated conflicts for the identification of real-time conflict-prone traffic conditions. Although vehicles in microsimulation do not collide, they have abundant interactions with each other and their motions are realistic because of the built-in car-following and lane-changing models. Furthermore, if proper attention to the correct calibration of the microsimulation model is given, traffic conditions before a traffic conflict can be used as a surrogate measurement to identify traffic collisions. Hence in this paper, simulated traffic conditions and the corresponding conflicts data will be utilised to estimate conflict-prone traffic conditions using machines learning classifiers (i.e. SVMs, *k*NN, RF).

B. Review of Real-Time Collision Classifiers

In order to reliably identify conflict-prone traffic conditions, potential classifiers need to be fast, accurate and suitable for real-time applications. Since there is no previous study concentrating on the identification of real-time conflict-prone conditions, the classifiers used to detect real-time traffic collisions were reviewed in order to choose the most appropriate techniques.

Real-time collision classifiers tend to relate real-time traffic measurements (coming usually from loop detectors) with the probability of a traffic collision. Early studies on real-time collision prediction (e.g. [26]–[28]) concentrated on analysing only traffic data prior to crash occurrences. Using relatively simple statistical techniques such as nonparametric Bayesian filters [26], loglinear modelling [27] and nonlinear canonical correlation analysis (NLCCA) [28] the aim of those studies was to estimate a relative collision or collision type risk probability given historical traffic and accident data. Although these studies accomplished a statistical relationship between real-time traffic and collision occurrence, they lack in terms of data sample size, classification accuracy and transferability issues and the implementation of their results was not suggested from future researchers [12], [29].

The state-of-the-art in real-time collision prediction modelling requires the utilization of data just before a collision occurrence (termed as collision-prone) as well as data of collision-free/normal traffic conditions. Both data categories describe traffic conditions on the road segment where a historical collision took place. Traffic data resembling collision-prone and normal traffic are usually employed in matched-case control study designs in which every collision-prone traffic condition is matched with a number of normal traffic cases in order to single out collision precursors (i.e. traffic indications of an imminent collision). Usually in matched-case control real-time collision prediction, the ratio of collision-prone to safe traffic conditions varies from 1:4 *(e.g.* [16]), 1:5 *(e.g.* [17], [18]) to 1:34 *(e.g* [4]). However the ratio between control and cases can prove essential for the classification results [13], [29]. Thus, the potential classifier needs to perform well without over-representing safe traffic conditions.

Methodologically, recent real-time collision prediction approaches are divided into logistic regression (e.g. [33], [34]) and artificial intelligence (AI)/machine learning approaches (e.g. [4], [6], [12], [35]–[37]).

With regards to logistic regression models, traditional logit [5] and Bayesian logistic regression [38] have been applied. However, regression models require the determination of critical odds ratio for the identification of collision-prone traffic conditions [39] and also rely heavily on distribution assumptions for both the collision frequency and the traffic parameters.

The first approaches within the machine learning domain for real-time collision prediction were concerned with Neural Network (NN) applications. For example, Pande and Abdel-Aty [12], [35] utilized three types of NNs (i.e. Probabilistic [12], Radial Basis Function [35] and Multilayer Perceptrons [35]) for real-time collision estimation in American freeways and demonstrated that NNs outperform statistical approaches without requiring distributional assumptions. Despite their learning and classification performance NNs usually require a large dataset for training [40]. However their major drawback is related to the incorporation of the "black-box" effect which prevents clear understanding of the model's underpinning properties [41]. Furthermore NN models often suffer from overfitting [36] and require extra computational resources to overcome it [40].

In order to deal with the drawbacks of regression models and NNs, Hossain and Muromachi proposed Bayesian Networks [4]. However, Bayesian Networks require a sufficiently large dataset to represent the probabilities of each of their nodes which make them difficult to implement with small and unbalanced datasets. Genetic Programming [37] was proposed by Xu et al to remove the "black-box" effect of machine learning approaches but their model faced difficulties with regards to transferability and practical implementation.

As a result, alternative classifiers were sought in this paper to overcome the main existing methodological drawbacks: i) the imbalance of collision/traffic datasets which over-represent safe traffic conditions compared to collision-prone, ii) the "black-box" effect and overfitting of NN models and iii) the inflexibility in the incorporation of correlated variables from regression models.

According to Dreisetl and Ohno-Machado [42] SVMs are flexible and have less over-fitting problems while *kNN* provides a case-based explanation on classification results, address the black-box problem and is easily transferrable because they do not require prior knowledge of any datasets. Additionally in a survey by Verikas *et al.* [43] it was demonstrated that RF are computationally light, have no overfitting problems, provide an insight on the importance of each predictor for the classification result and perform better or similarly to other classifiers such as SVMs for a large number of applications (e.g. cancer detection, face recognition and network intrusion detection). In the same spirit, Rokach [44] concluded that RF can handle a large number of predictors without being computationally heavy. The aforementioned advantages of the three algorithms (i.e. SVMs, kNN and RF) indicate that these algorithms could be potential conflict detection classifiers and were reviewed for their applicability.

C. Applications of SVMs, kNN and RF in Real-Time Classification Problems

SVM models have been applied to real-time collision and traffic flow predictions. For instance, Li *et al.* [45] compared the findings from the SVMs with that of the popular negative binomial models in predicting motorway collisions. Their results showed that SVM models have a better goodness-of-fit in comparison with negative binomial models. Their findings were in line with the study of Yu and Abdel-Aty [36] who compared the results from the SVM and Bayesian logistic regression models for evaluating real-time collision risk demonstrating the better goodness-of-fit of the SVM models. The prediction of side-swipe accidents using SVMs was evaluated in Wang *et al.* [46] by comparing SVMs with Multilayer Perceptron Neural Networks. Both techniques showed similar accuracy but SVMs led to better collision identification at higher false alarm rates. More recently, Dong *et al.* [47] demonstrated the capability of SVMs to assess spatial proximity effects for regional collision prediction.

On the other hand, kNN has recently been applied in the area of short-term traffic prediction due to the fact that it is one of the simplest data mining algorithms. Zhang *et al.* [34] made a first attempt to use kNN for traffic flow prediction using occupancy rate, vehicle speed and weather data. They compared the results from kNN with the results of backpropagation Neural Networks and showed than kNN classification was more accurate and transferable. Furthermore, Hou *et al.* [48] argued that although kNN has a relatively slow computing speed, it is suitable for real-time applications. Lastly, in comparison with SVM, kNN have better transferability as suggested by Zhang *et al.* A recent study on variable selection for real-time collision prediction [49] utilized kNN and showed that kNN can produce efficient collision predictions when utilized with traffic data aggregated in 5-minute and 10-minute intervals.

RF has mainly been applied in the area of real-time collision prediction for variable selection purposes. Its purpose within real-time collision prediction was to select the most important variables to be used in the subsequent modelling. Abdel-Aty *et al.* [31] initially combined RF for variable selection with Neural Networks and suggested that the resulting classifiers can efficiently distinguish collision-prone traffic conditions. Improved classification results were also demonstrated when RF were combined with logistic regression [11] and genetic programming [37] in order to identify important variables to be used in real-time collision models. To the author's knowledge, however, there is no study employing RF for distinguishing between collision-prone and safe traffic conditions.

The recent work on SVMs proves that they are an efficient classifier as well as a successful predictor when applied to traffic collisions prediction. Hence, it is a potential candidate for detecting conflict-prone conditions effectively. Moreover, the simplicity of *k*NN and its real-time applicability as suggested by studies on real-time traffic prediction provides an alternative algorithm that can be used for classifying traffic conditions. Lastly, the effectiveness of RF and its succesful application on other domains as well as its primary use as a variable selection method as suggested by studies on real-time collision prediction provides an alternative algorithm that can be used in addition to SVMs and kNN for detecting conflict-prone conditions in real-time.

D. Literature Review Findings

In summary, it can be concluded that data from a traffic micro-simulation tool (e.g. VISSIM) and relevant traffic conflicts from the SSAM have the potential to improve real-time highway safety assessment. Although vehicles in micro-simulation do not collide, they have abundant interactions with each other and their motions are realistic because of the built-in car-following models. Consequently, if proper attention to the correct calibration of the micro-simulation model is given, traffic conditions before a traffic conflict can be used as a surrogate measurement to identify traffic collisions. However, existing studies utilising VISSIM/SSAM concentrate on the investigation of the correlation between traffic collisions and traffic conflicts so as to estimate the number of collisions and the impact of interventions through the use of traffic conflicts. In this paper, simulated traffic conditions and the corresponding conflicts data are utilised to estimate conflict-prone traffic conditions in real-time by the use of machines learning classifiers (i.e. SVMs *k*NN and RF).

Another issue that needs further investigation by the application of classification algorithms is the temporal aggregation of traffic data. Previous work on segment-based collision prediction indicated that 5-10-minute aggregated traffic data (e.g. [5]) offer an ideal balance between capturing the microscopic traffic fluctuations and enabling sufficient time to traffic authorities for introducing interventions. Such temporal aggregation may not be optimal for the case of (semi)autonomous vehicles which need a reliable prediction of unsafe traffic conditions as fast as possible. Therefore, different temporal aggregation intervals (i.e. *30*-second, *1*-minute, *3*-minute and *5*-minute) will be tested in order to identify the aggregation offering the best results in real-time conflict-prone traffic conditions estimation.

III. Description of Classification Algorithms

The objective of this study is to identify conflict-prone traffic conditions from highly disaggregated data by using the SVM, kNN and RF classifiers.

SVMs belong to the larger group of supervised learning algorithms and kernel methods. In supervised learning, there exists a set of example input vectors $\{\mathrm{x_n}\}_{\mathrm{n=1}}^{\mathrm{N}}$ along with corresponding targets $\{\mathrm{t_n}\}_{\mathrm{n=1}}^{\mathrm{N}}$, the latter of which corresponds to class labels. In this study, the two classes are defined as '*dangerous*' when t=1 and '*safe*' when t=0. The purpose of learning is to acquire a model of how the targets rely on the inputs and use this model to classify or predict accurately future and previously unseen values of x.

An SVM classifier is based on the following functional form:

$$y = f(x; w) = \sum_{i=1}^{N} w_i K(x, x_i) + w_0 = w^T \varphi(x) \qquad (1)$$

In (1), K (x,x$_\mathrm{i}$) is a kernel function, which defines a basis function for each data point in the training dataset, w_i are the weights (or adjustable parameters) for each point, and w_0 is the constant parameter. The output of the function is a sum of M basis functions $\varphi(\mathrm{x}) = [\varphi_1(\mathrm{x}), \varphi_2(\mathrm{x}), \ldots c, \varphi_M(\mathrm{x})])$ which is linearly weighted by the parameters w.

SVM, through its target function, tries to find a separating hyper-plane to minimize the error of misclassification while at the same time maximize the distance between the two classes [36]. The produced model is sparse and relies only on the kernel functions associated with the training data points which lie either on the margin or on the wrong side. These data points are referred to as "Support Vectors" (SVs).

*k*NN is a non-parametric learning algorithm which is simple but effective in many cases [50]. For a data record t to be classified its k nearest neighbours are retrieved and this forms a neighbourhood of t. During training, each t is assigned to a class if the majority of the k neighbours of t belong to this particular class. However, an appropriate value for k is needed to apply a *k*NN approach and the success of classification is very much dependent on this value [51].

RF belongs to the group of ensemble classifiers and more specifically to the group of bagging algorithms. Bagging algorithms make use of only one learning algorithm and modify the training set by using the bagging algorithm to create new training sets [52]. RF is an evolution of bagged trees and uses the bagging algorithm along with the random subspace method proposed by Ho [53]. Each tree is built using the impurity Gini index [54]. Nevertheless, only a random subset of the input features is used for the construction of the tree and no pruning takes place. For each new training dataset, one-third of the samples is randomly neglected and forms the out-of-bag (OOB) samples. The samples that are not neglected are used for building the tree. For every constructed tree the OOB samples are used as a validation dataset and the misclassification OOB error is estimated. When a new data record (say t) needs to be classified, it is run through all the constructed trees and a classification result for every tree is obtained. The majority vote over all the classification results from all the constructed tree is chosen as the classified label for that specific data record [43]. However, an appropriate value for the number of features used for splitting a node of a tree needs to be tuned by the user in order for the OOB misclassification error to be as low as possible [43].

IV. Data Description and Processing

This study aims to examine the effectiveness of SVM, *k*NN and RF classifiers in identifying conflict-prone traffic conditions using data from a traffic micro-simulation (i.e. VISSIM) and the SSAM. As discussed in the literature review section, the fundamental issue relating to this approach was the building and calibrating of a traffic micro-simulation model using real-world traffic data. Fig. 1 shows the overall methodology of capturing the required data for classification purposes and how the results from the classification algorithms can be applied to different real-time applications.

As can be seen from Fig.1, link-level disaggregated traffic data from loop detectors and GPS-based probe vehicles were obtained from the UK Highways England Journey Time Database (JTDB). Link-level data corresponded to every day of the years 2012 and 2013 and included average travel speed and volume at 15-minute intervals. It should be noted here that 15-minute traffic data corresponded to link-based average speed, volume and journey time of all vehicles between two junctions.

A 4.52-km section of M62 (a motorway in England) between junction 25 and 26 was selected as the study area. The segment has three lanes in each direction. On-ramp and off-ramp traffic were not taken into account because relevant data were not available. In order to build a robust micro-simulation model, the JTDB data were split into four scenarios for each year:

- Morning peak hours (06:00 – 09:30)
- Morning off-peak hours (09:30-13:00)
- Afternoon off-peak hours (13:00-15:45)
- Afternoon peak hours (15:45-19:15)

For each of these scenarios the 15-minute traffic volumes and the cumulative speed distribution of the roadway segment were extracted and employed as input to VISSIM. Furthermore, the vehicle composition for 2012 and 2013 was also obtained from the UK Department of Transport [55] and was used to build a micro-simulation model. The road segment was manually coded in VISSIM using a background image from OpenStreetMap [56]. The allocation of data

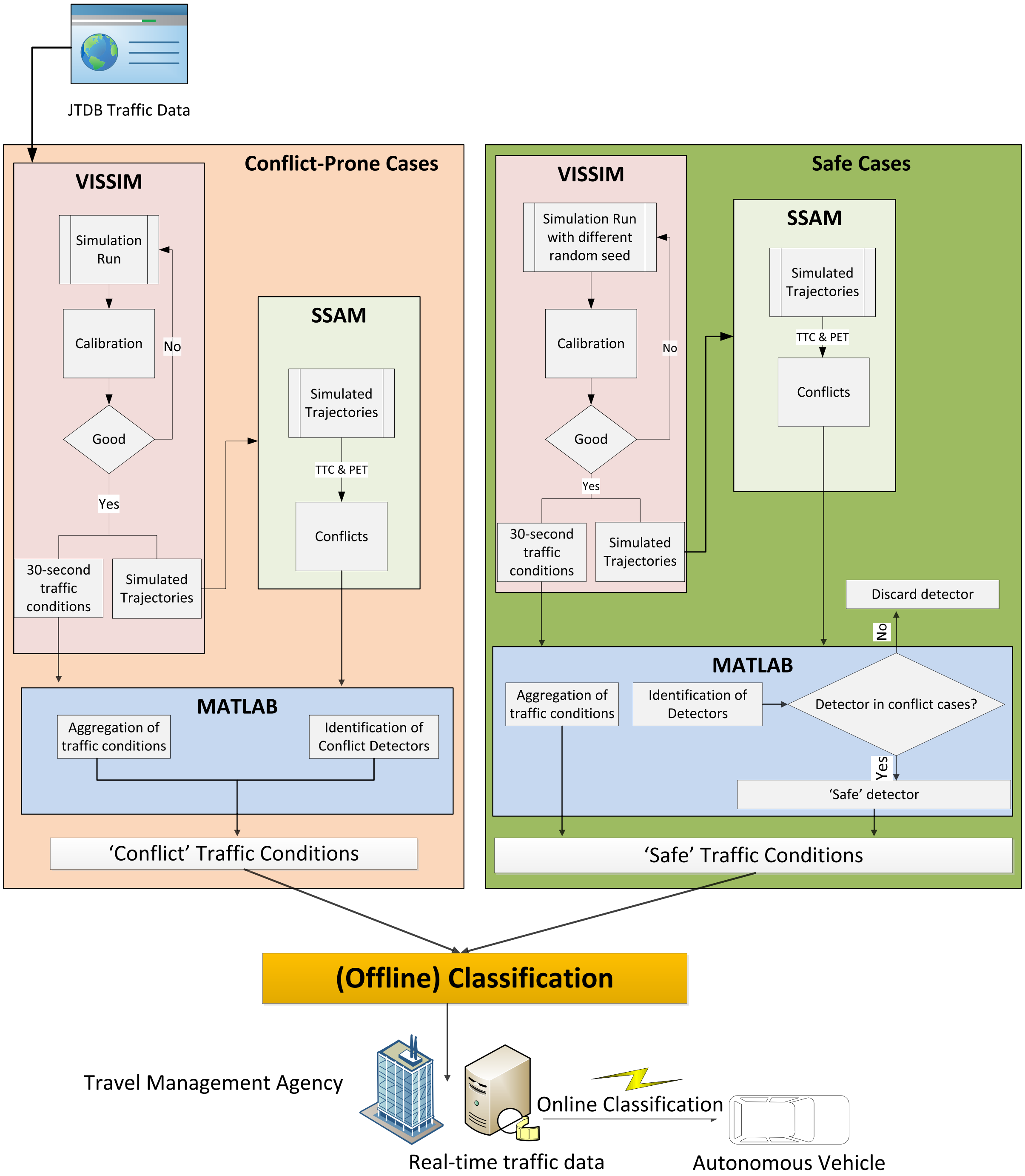

Fig. 1. Flow chart of the procedure followed to classify traffic conditions for real-time conflict detection from AVs.

collection detectors for the acquisition of traffic data was decided to be 300m in order to resemble the spacing of detectors in previous studies on real-time collision prediction on motorways (e.g. [4], [36]). Such studies also investigated real-time safety assessment using traffic conditions measured from detectors with such spacing and claimed that the prediction can be performed in real-time. Therefore, a similar spacing was used in the micro-simulation model in this paper

In order for the micro-simulation to be initiated, the car-following model needed to be defined in VISSIM, the Wiedemann 99 model was selected because it applies to motorway scenarios [14]. The Wiedemann model is characterised mainly by the three parameters in VISSIM; the standstill distance, the headway time and the following variation [14]. The standstill distance describes the average standstill distance between two vehicles. The headway time is the time gap (in seconds) which a driver wants to maintain at a certain speed. On the other hand, the following variation defines the desired safety distance a driver allows before moving closer to a vehicle in front.

According to the guidelines from the Federal Highway Administration (FHWA) [57], the GEH-statistic [58] and the

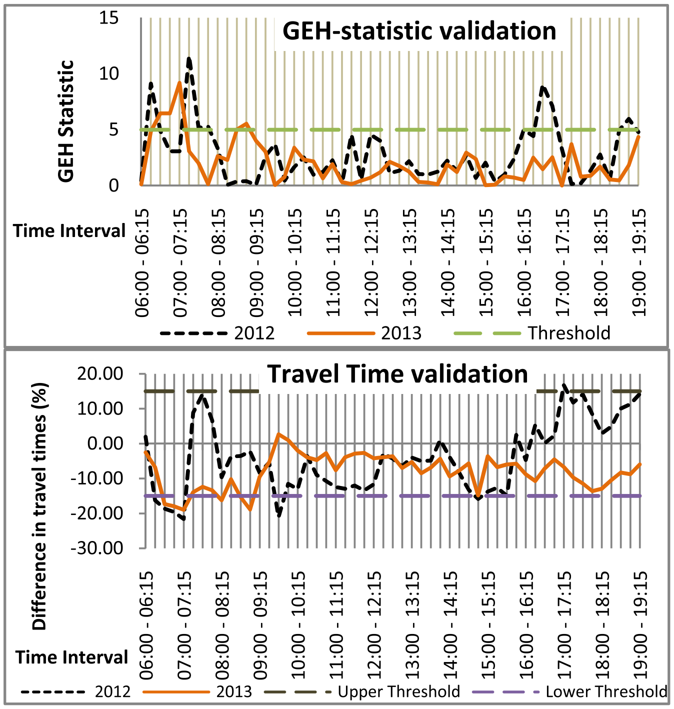

Fig. 2. GEH statistic and Travel time validation for each time interval and year.

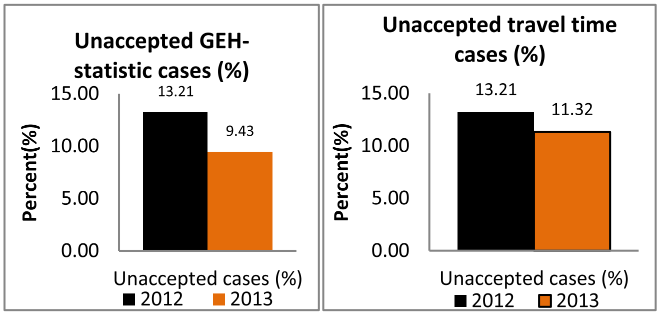

Fig. 3. Percentage of unaccepted cases for each year regarding the GEH statistic and travel time.

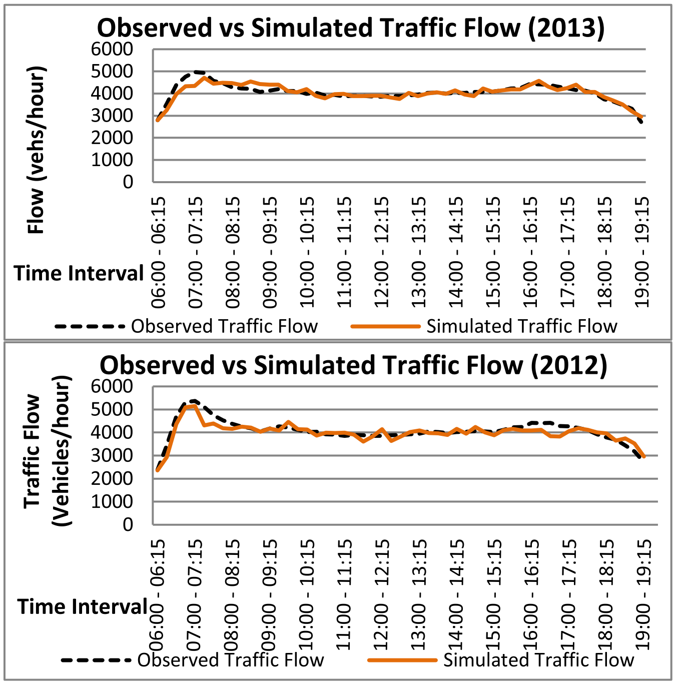

Fig. 4. Observed vs Simulated Traffic flow for each year.

link travel time were used. The GEH statistic correlates the observed traffic volumes with the simulated volumes as shown below:

$$GEH = \sqrt{\frac{(V_{sim} - V_{obs})^2}{\frac{V_{sim} + V_{obs}}{2}}} \tag{2}$$

where V_{sim} is the simulated traffic volume and V_{obs} is the observed traffic volume.

After a number of trial simulations, the best GEH values were obtained by using the following parameters for the Wiedemann 99 car following model:

- Standstill distance: 1.5 m
- Headway time: 0.9 sec
- Following variation: 4 m

For the simulation to efficiently resemble real-world traffic it is essential that [57]:

1) GEH statistic <5 for more than the 85% of the cases
2) The differences between observed and simulated travel times is equal or below 15% for more than 85% of the simulated cases.

The validation results are summarized in Fig. 2 and 3, and the comparison between traffic flow and travel time in simulation and reality are depicted in Fig. 4 and 5. The calibration was performed using the entire simulated dataset (from all four periods) and the observed traffic conditions and conflicts so as to have a unified dataset.

In the simulations that were undertaken, the GEH values for most of the time intervals were found to be less than five. However, there were intervals where GEH values were found to be between 5 and 10. According to [59] these values indicated either a calibration problem or a data problem. Because of the large number of simulations undertaken (∼1000 for every scenario) it was assumed that the bad GEH values related to the highly aggregated traffic data (i.e. 15-minute by road-level). Therefore, it was decided to keep the simulation results for the intervals with GEH values between 5 and 10.

After calibrating the simulations, three additional simulations with different random seeds were run resulting in a total of four different simulation results for each of the scenarios. The number of additional runs was chosen in order to cope with the imbalance between conflict and safe conditions which can prove essential for classification purposes [13]. The four different simulations were used for the matched-case control structure, where the first simulation was used to acquire the traffic conflicts and the other three were used to resemble the normal traffic conditions.

For the extraction of traffic conflicts, the vehicle trajectory files exported from VISSIM were inserted to the SSAM. Conflicts were detected if the TTC value between two vehicles was below 1.5 seconds and the PET value was below 4 seconds which are the default values used in SSAM [15]. Only lane-changing and rear-end conflicts were considered

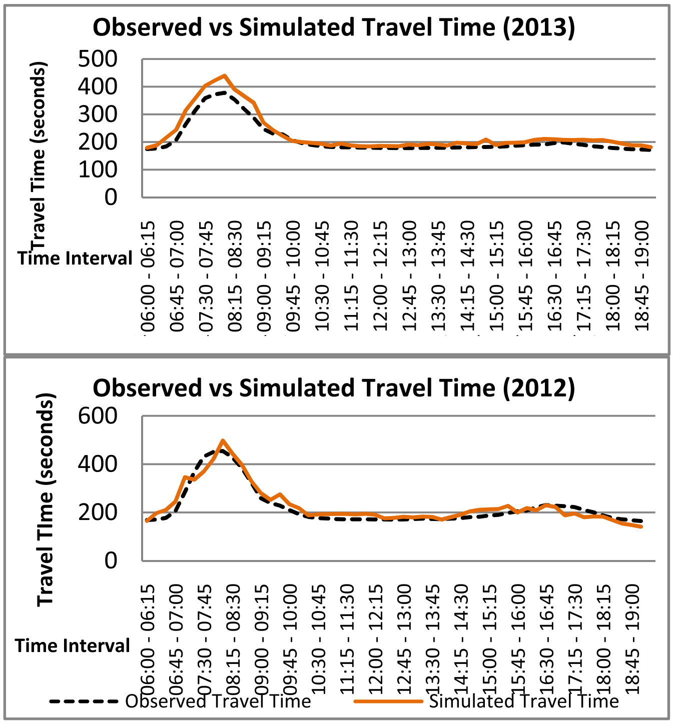

Fig. 5. Observed vs Simulated travel time for each year.

TABLE I
TRAFFIC CONFLICTS STATISTICS

Conflict type	Count	Average TTC (sec)	Average PET (sec)	Maximum deceleration (m/s^2)
lane change	2199	0.9561	0.2065	-5.7237
rear end	876	0.9614	0.2122	-5.4266

according to the SSAM manual for motorway scenarios. A total of 3,075 traffic conflicts and the corresponding 9,225 conflict-prone traffic conditions were gathered for further analysis. Table I presents the type, count and average statistics of the conflicts used for the analysis.

In order for the conflicts to be validated, the Crash Potential Index (CPI) was used as suggested by Cunto [60]. CPI is calculated through the following equation:

$$CPI_i = \frac{\sum_{t=t_{i_i}}^{tf_i} (P(MADR^{(a_1,a_2,\ldots,a_n)} \leq DRAC_{i,t}) \cdot \Delta\text{t} \cdot b}{T_i} \qquad (3)$$

where CPI_i is the CPI for vehicle i, $DRAC_{i,t}$ is the deceleration rate to avoid the crash (m/s^2), $MADR^{(a_1,a_2,\ldots,a_n)}$ is a random variable following normal distribution for a given set of environmental attributes, t_{i_i} and t_{f_i} are the initial and final simulated time intervals for vehicle i, Δt is the simulation time interval (sec), T_i is the total travel time for vehicle i and b is a binary state variable denoting a vehicle interaction. For MADR according to [60] a normal distribution with average of 8.45 for cars and 5.01 for HGVs with a standard deviation of 1.4 was assumed for daylight and dry pavements. The results for the calibration of the conflicts are shown in Fig.6

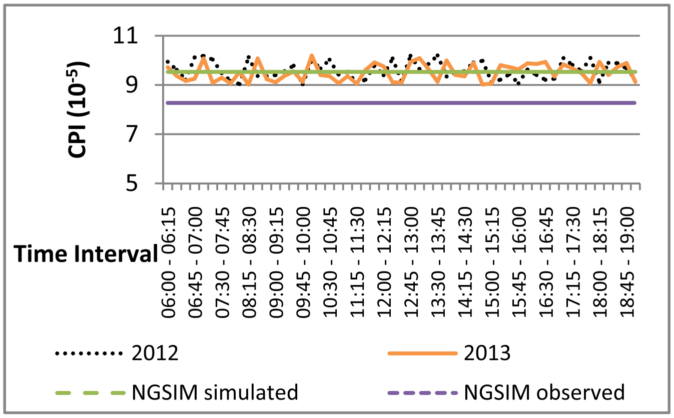

Fig. 6. Conflicts validation.

In Fig. 6 it is shown that for the majority of the time intervals, CPI is similar to the simulated CPI of the NGSIM dataset and close to the values of the observed NGSIM CPI. Therefore, it can be concluded that the simulated conflicts resembled realistic hazardous scenarios.

In the last step of the data processing, a MATLAB [61] code was developed in order to match the conflicts (exported from the SSAM) with the traffic conditions (acquired from VISSIM). The estimated conflicts were filtered again to obtain conflicts with TTC below 1.3 seconds and PET below 1 second in order to obtain conflicts which are difficult to avoid. That is because TTC below 1.3 seconds is lower than the average human reaction time [62] and PET values close to zero show imminent collisions [15]. Other TTC values were also tested however the value 1.3 provided a sample containing sufficient cases of near-misses (TTC<0.5seconds) and conflicts with TTC close to the human reaction time (TTC<1.5seconds). As conflicts extracted by SSAM and traffic conditions acquired by VISSIM were time stamped it was concluded that the issue of incorrectly reported collision times has largely been overcome.

For each of the conflicts, the nearest upstream detector on the road segment was identified by comparing the time of the conflict with the time that the vehicles passed each detector. This specific detector was marked as "conflict detector". Traffic data were extracted for every conflict detector, the corresponding upstream and downstream detectors on the same lane and the detector in the adjacent lane for every time interval. The traffic measurements for these detectors were marked as "conflicts" because they represent the traffic conditions near the time when the conflict occurred. In order to obtain the non-conflict cases in a matched-case control design, for every conflict detector the conflicts for the other three simulation runs were checked to see if any conflicts happened near them in these runs. If there was no conflict, the traffic measurements from that detector were obtained to represent safe conditions. Otherwise the detector was discarded. For each of the detectors and for every time interval the number of vehicles, the vehicle speeds and the vehicle accelerations were extracted.

The ultimate purpose of the approach, as seen in Fig.1, is that after the offline classification, traffic management agencies could use the classification algorithms with real-time

TABLE II
CLASSIFICATION ACCURACY PER DATA AGGREGATION INTERVAL (SAMPLE SIZE 12300 CASES WITH 10-FOLD CROSS-VALIDATION)

CLASSIFIER	ACCURACY			
	30-SECONDS	1-MINUTE	3-MINUTE	5-MINUTE
SVM	75.60%	76.40%	77.50%	78.40%
35-NN	75.10%	75.50%	76.30%	77.80%
RF	76.30%	76.95%	78.49%	79.44%

traffic data to inform AVs about conflict-prone traffic conditions Similar to previous real-time safety assessment studies [5] when a conflict is detected the information could be broadcasted to drivers and AVs so that the vehicles adjust to a lower speed, thus reducing the probability of a conflict.

V. RESULTS AND DISCUSSION

Classification methods of SVM, *k*NN and RF have been applied to a unified dataset containing all the cases ("*conflicts*" and "*safe*") as discussed above. SVM and *k*NN were developed using the Statistics and Machine Learning Toolbox of MATLAB and RF were developed using WEKA [63].

SVMs depend on the kernel functions to perform the classification. The most popular kernels used in the SVM classification are the linear, polynomial and Gaussian or radial basis function (RBF). In this study, the Gaussian kernel has been used because existing research suggests that it provides more accurate results [36]. The Gaussian kernel is calculated through the equation:

$$K\left(x_i, x_j\right) = exp\left(-\gamma \left\|x_i - x_j\right\|^2\right) \tag{4}$$

where γ determines the width of the basis function. The coefficient γ was set to 0.5 because the targets of the classification lie in the interval {0,1}.

The *k*NN classifier, on the other hand, requires tuning the important parameter - the number of nearest neighbours (*k*). A usual approach is to perform tests with different *k* values starting from 1 and ending at the square root of the number of observations [64]. In this paper after a grid search among different k values, the best results came from using $k = 35$.

The RF classifier, on the other hand, requires tuning the number of features used for node splitting. A usual approach as suggested by Breiman [54] is to use the value $\log_2(\mathrm{PR} + 1)$, where PR is the number of predictors used for classification. This approach was used in this paper for building the RF classifier. A hundred trees (i.e. the default value in WEKA) were used to make an ensemble.

To test the performance of the three different algorithms (i.e. SVM, *35*-NN and RF) the classification accuracy was initially tested for each of the temporal aggregation intervals. The results are summarised in Table II.

Likewise, existing studies on collision-prone traffic conditions estimation, classification performance increases with higher temporal aggregation. Thus, traffic data aggregated in *5*-minute time intervals have proved to be a better conflict precursor than any other temporal aggregation used in this study. This is probably related to the noise inherent to *30*-second and *1*-minute aggregated data and lack of *3*-minute data to capture accurately the traffic dynamics leading to a conflict.

To further investigate the performance of the classifiers for real-time conflict-prone traffic conditions identification as well as to cope with the imbalance of the dataset (because *conflict* to *safe* conditions ratio is 1:3) several metrics were employed to evaluate the performance of the classifiers. These metrics are sensitivity, specificity, precision, *G*-means and *F*-measure and are defined in (*5*) - (*9*) according to [6]:

$$\text{Sensitivity} = \frac{\mathrm{T_{conflict}}}{\mathrm{T_{conflict}} + \mathrm{F_{safe}}} \tag{5}$$

$$\text{Specificity} = \frac{\mathrm{T_{safe}}}{\mathrm{T_{safe}} + \mathrm{F_{conflict}}} \tag{6}$$

$$\text{Precision} = \frac{\mathrm{T_{conflict}}}{\mathrm{T_{conflict}} + \mathrm{F_{conflict}}} \tag{7}$$

$$\text{G-means} = \sqrt{Sensitivity * Specificity} \tag{8}$$

$$\text{F-measure} = \frac{2 * precision\ *\ sensitivity}{precision + sensitivity} \tag{9}$$

where $T_{conflict}$ represents a correct detection of conflict-prone traffic conditions identified as *conflict-prone*, $F_{conflict}$ represents an incorrect detection of conflict-prone traffic conditions identified as *safe*, T_{safe} is a safe traffic condition instance correctly identified as *safe* and F_{safe} is a safe traffic condition instance falsely identified as *conflict-prone*.

The sensitivity statistic shows the correct classification accuracy with respect to conflict-prone traffic conditions, while the specificity statistic shows the classification accuracy in terms of safe conditions. Precision is used for identifying the classification accuracy among each class. G-means is used to check whether the use of an imbalance dataset (1:3; conflicts vs safe) has any negative impact on the balanced qualification accuracy. Lastly, the *F*-measure is a metric which resembles the conflict-prone classification ability of the classifier models. Results for all the above-mentioned performance metrics for the classifiers are summarised in Table. III.

From Table III it can also be observed that RF demonstrates a higher *sensitivity* and *specificity* compared to SVMs and *k*NN. This implies smaller Type I and Type II errors because both conflict-prone and safe conditions have a better chance of being correctly classified, especially when using *30*-second traffic data. The performance of all classifiers regarding *sensitivity* and *specificity* improves with higher temporal aggregation reaching its best value when *5*-minute traffic data are classified. On the other hand, the low scores of *sensitivity* imply that conflict-prone conditions are not accurately classified. This is most probably due to the class-imbalance problem which is further resembled on the relatively high *precision* scores. High *precision* but low *sensitivity* is an indicator that the classifiers perform well in classifying traffic conditions but most of the correct classifications correspond to safe traffic conditions (which form most of the sample). The *F-Measure* results of the classifiers in Table. II enhance the observation that conflicts are difficult to be detected by all

TABLE III

Classification Performance Metrics Per Data Aggregation Interval (Sample Size 12300 Cases With 10-Fold Cross-Validation)

	Sensitivity	Specificity	Precision	G-means	F-Measure
30-second Data					
SVM	0.059	0.759	0.637	0.211	0.108
35-NN	0.023	0.753	0.551	0.131	0.044
RF	0.081	0.764	0.735	0.249	0.146
1-minute Data					
SVM	0.091	0.765	0.717	0.264	0.161
35-NN	0.042	0.757	0.657	0.179	0.079
RF	0.119	0.771	0.744	0.303	0.205
3-minute Data					
SVM	0.138	0.775	0.781	0.327	0.234
35-NN	0.085	0.764	0.716	0.255	0.152
RF	0.175	0.782	0.833	0.370	0.289
5-minute Data					
SVM	0.179	0.783	0.813	0.375	0.294
35-NN	0.159	0.778	0.765	0.352	0.263
RF	0.208	0.790	0.871	0.406	0.336

three algorithms probably due to the class imbalance problem, as well as the noise included in lower temporal aggregation intervals.

Finally, observing the *G-means* metric results, which show the balanced classification ability of the classifiers, RFs outperform the other algorithms. The value of the G-means metric increases in bigger temporal aggregation intervals but still is in correspondence with the class imbalance problem.

The classification accuracy obtained from each of the three classifiers agrees with the results from existing studies (e.g. [6], [27] which use actual collision data and more precise traffic data. The results of the classifiers regarding accuracy, *G*-Means and *F*-measure are comparable to the findings by Sun and Sun [6] who employed a Dynamic Bayesian Network (DBN) classifier.

Their DBN classifier achieved an overall accuracy of 76.6% which is similar to the accuracy of most of the conflict-based classifiers in this paper. It should also be observed that using *3*-minute and 5-minute traffic data with SVMs and RFs, the accuracy performance of the developed classifiers increases to 78 -79%. Furthermore, it is shown that even though the kNN classifier is considered to be a simple algorithm, its performance is similar when using *3*-minute and 5-minute traffic data. Regarding *G*-means and F-measure the DBN classifier in [6] has a value of 0.76 and 0.51 respectively which is superior to the classifiers in the current study data. This shows that further research is needed to overcome the data imbalance for better detection of conflict-safe traffic conditions.

In summary, it can be concluded that traffic data aggregated in a *5*-minute interval have proved to be the best temporal aggregation in classifying conflict-prone traffic conditions. However, it is noted that the achieved accuracy and sensitivity are relatively low. A reason behind these low metrics might be that traffic conditions leading to a conflict might not differ so significantly from normal traffic conditions as collision-prone traffic conditions do. Improvements regarding the data imbalance problem need to be made to increase the G-means and F-measure metrics for the classifiers.

VI. Conclusions

This paper developed a simulation approach to detect traffic conflict-prone traffic conditions in real-time. This approach overcame two issues associated with the classification of collision-prone traffic conditions employed in existing studies: (i) the temporal traffic data aggregation problem and (ii) the issues surrounding the incorrect reporting of collision time and the corresponding misrepresentative pre-collision traffic conditions. Furthermore, in this paper the Random Forest algorithm was applied for classification and not for the task of variable selection as in previous literature regarding real-time collision prediction. Significant efforts were devoted to calibrating the traffic simulation model in VISSIM.

The classification results showed that traffic micro-simulation along with safety thresholds to detect conflicts from the SSAM model could be used in real-time safety assessment. The accuracy of the SVM, *k*NN and RF classifiers was found to be in-line with recent studies on real-time collision prediction which used actual collision data along with the corresponding traffic data. The superiority of *5*-minute temporal aggregation in the classification results comes in agreement with safety experts who utilized *5*-minute aggregated data to understand the traffic fluctuations and the occurrence of traffic collisions. Thus, having overcome the misreported collision time simulation-based data can better represent traffic conditions before the occurrence of a dangerous vehicle encounter. In terms of practical applications in real-time, a library of calibrated functions can be developed offline (for example, one for lane closures) and used them as an online application. In the same principle of real-time collision prediction, traffic management agencies could utilize real-time conflict prediction and warn road users if conflict-prone conditions are present. Since the mechanism leading to a conflict and the mechanism leading to collision present similarities, the correct real-time identification of conflict-prone conditions would lead to safer real-time traffic because collisions are a fraction of the observed conflicts.

Researchers should however be cautious if highly disaggregated traffic data (i.e. 30-second) are utilized in estimating real-time conflicts for risk assessment especially in applications of advanced driver-assistance systems (ADAS) and autonomous vehicles (AVs).Special attention shall be given in the validation using detailed real-world data e.g. video surveillance data or radar-based data for conflicts and lane-based traffic data under many incident conditions (e.g. temporary work zones or lane closures). Further research shall be devoted to solving the issue with the data imbalance as identified in this study by the low G-means and F-measure metrics. Since ‘ground truth’ data for conflicts were not available in this study, it would be beneficial to identify the most effective TTC threshold so as to enhance model predictive performance. Hence, TTC and PET metrics which consider all kinds of conflicts as well as the acceleration of vehicles should also be researched. Finally, sensitivity analysis for choosing the cases:controls ratio and feature selection for the calibration of the classifiers may enhance the classification results.

Acknowledgment

The authors take full responsibility for the content of this paper and any errors or omissions. Research data for this paper are available on request from Christos Katrakazas (c.katrakazas@lboro.ac.uk).

References

[1] S. Roshandel, Z. Zheng, and S. Washington, "Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis," *Accident Anal. Prevention*, vol. 79, pp. 198–211, Jun. 2015.

[2] T. Hummel, M. Kühn, J. Bende, and A. Lang, "Advanced driver assistance systems: An investigation of their potential safety benefits based on an analysis of insurance claims in Germany," Germany Insurance Assoc., London, U.K., Tech. Rep. FS 03, 2011.

[3] S. Thrun, "Toward robotic cars," *Commun. ACM*, vol. 53, no. 4, pp. 99–106, 2010.

[4] M. Hossain and Y. Muromachi, "A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways," *Accident Anal. Prevention*, vol. 45, pp. 81–373, Mar. 2012.

[5] M. Abdel-Aty and A. Pande, "The viability of real-time prediction and prevention of traffic accidents," in *Efficient Transportation and Pavement Systems*, I. L. Al-Qadi, T. Sayed, N. Alnuaimi, and E. Masad, Eds. London, U.K.: Taylor & Francis, 2005, pp. 215–226.

[6] J. Sun and J. Sun, "A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data," *Transp. Res. C, Emerg. Technol.*, vol. 54, pp. 176–186, May 2015.

[7] M.-I. Imprialou, "Developing accident-speed relationships using a new modelling approach," Ph.D. dissertation, School Civil Building Eng., Loughborough Univ., Loughborough, U.K., 2015.

[8] K. El-Basyouny and T. Sayed, "Safety performance functions using traffic conflicts," *Safety Sci.*, vol. 51, no. 1, pp. 160–164, 2013.

[9] U. Shahdah, F. Saccomanno, and B. Persaud, "Application of traffic microsimulation for evaluating safety performance of urban signalized intersections," *Transp. Res. C, Emerg. Technol.*, vol. 60, pp. 96–104, Nov. 2015.

[10] F. Amundsen and C. Hyden, in *Proc. 1st Workshop Traffic Conflicts*, 1977, p. 87.

[11] H. M. Hassan and M. A. Abdel-Aty, "Predicting reduced visibility related crashes on freeways using real-time traffic flow data," *J. Safety Res.*, vol. 45, pp. 29–36, Jun. 2013.

[12] M. Abdel-Aty and A. Pande, "Identifying crash propensity using specific traffic speed conditions," *J. Safety Res.*, vol. 36, no. 1, pp. 97–108, Jan. 2005.

[13] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[14] *PTV VISSIM 6 User Manual*, PTV AG, Karlsruhe, Germany, 2013.

[15] L. Pu and R. Joshi, "Surrogate safety assessment model (SSAM): Software user manual," Federal Highway Admin., Washington, DC, USA, Tech. Rep. FHWA-HRT08-050, 2008, p. 96.

[16] P. Bonsall, R. Liu, and W. Young, "Modelling safety-related driving behaviour—Impact of parameter values," *Transp. Res. A, Policy Pract.*, vol. 39, no. 5, pp. 425–444, 2005.

[17] F. Huguenin, A. Torday, and A. Dumont, "Evaluation of traffic safety using microsimulation," in *Proc. 5th Swiss Transp. Res. Conf.*, 2005, p. 1.

[18] M. M. Minderhoud and P. H. L. Bovy, "Extended time-to-collision measures for road traffic safety assessment," *Accident Anal. Prevention*, vol. 33, no. 1, pp. 89–97, Jan. 2001.

[19] J. Archer, "Indicators for traffic safety assessment and prediction and their application in micro-simulation modelling: A study of urban and suburban intersections," M.S. thesis, Dept. Infrastruct., Division Transp. Logistics, Royal Inst. Technol., Stockholm, Sweden, 2005.

[20] D. Gettman and L. Head, "Surrogate safety measures from traffic simulation models," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1840, pp. 104–115, Jan. 2003.

[21] F. Huang, P. Liu, H. Yu, and W. Wang, "Identifying if VISSIM simulation model and SSAM provide reasonable estimates for field measured traffic conflicts at signalized intersections," *Accident Anal. Prevention*, vol. 50, pp. 1014–1024, Jan. 2013.

[22] W. Young, A. Sobhani, M. G. Lenne, and M. Sarvi, "Simulation of safety: A review of the state of the art in road safety simulation modelling," *Accident Anal. Prevention*, vol. 66, pp. 89–103, May 2014.

[23] M. Essa and T. Sayed, "Simulated traffic conflicts do they accurately represent field-measured conflicts?" *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2514, pp. 48–57, Nov. 2015.

[24] R. Fan, H. Yu, P. Liu, and W. Wang, "Using VISSIM simulation model and Surrogate Safety Assessment Model for estimating field measured traffic conflicts at freeway merge areas," *IET Intell. Transp. Syst.*, vol. 7, no. 1, pp. 68–77, Mar. 2013.

[25] M. Essa and T. Sayed, "Transferability of calibrated microsimulation model parameters for safety assessment using simulated conflicts," *Accident Anal. Prevention*, vol. 84, pp. 41–53, Nov. 2015.

[26] C. Oh, J.-S. Oh, S. G. Ritchie, and M. Chang, "Real-time estimation of freeway accident likelihood," Dept. Civil Environ. Eng. Inst. Transp. Stud., Univ. California, Irvine, CA, USA, Tech. Rep. UCI-ITS-TS-WP-00-8, 2001.

[27] C. Lee, F. Saccomanno, and B. Hellinga, "Analysis of crash precursors on instrumented freeways," *Transp. Res. Rec.Transp. Res. Rec., J. Transp. Res. Board*, vol. 1784, no. 1, pp. 1–8, 2002.

[28] T. F. Golob and W. W. Recker, "A method for relating type of crash to traffic flow characteristics on urban freeways," *Transp. Res. A, Policy Pract.*, vol. 38, no. 1, pp. 53–80, 2004.

[29] C. Xu, P. Liu, and W. Wang, "Evaluation of the predictability of real-time crash risk models," *Accident Anal. Prevention*, vol. 94, pp. 207–215, Sep. 2016.

[30] M. Ahmed and M. Abdel-Aty, "A data fusion framework for real-time risk assessment on freeways," *Transp. Res. C, Emerg. Technol.*, vol. 26, pp. 203–213, Jan. 2013.

[31] M. Abdel-Aty, A. Pande, A. Das, and W. J. Knibbe, "Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2083, no. 2083, pp. 153–161, 2008.

[32] M. M. Ahmed and M. Abdel-Aty, "The viability of using automatic vehicle identification data for real-time crash prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 459–468, Jun. 2012.

[33] M. Abdel-Aty, N. Uddin, A. Pande, F. Abdalla, and L. Hsia, "Predicting freeway crashes from loop detector data by matched case-control logistic regression," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1897, pp. 88–95, Jan. 2004.

[34] C. Xu, W. Wang, P. Liu, and F. Zhang, "Development of a real-time crash risk prediction model incorporating the various crash mechanisms across different traffic states," *Traffic Injury Prevention*, vol. 16, no. 1, pp. 28–35, Jan. 2014.

[35] A. Pande and M. Abdel-Aty, "Assessment of freeway traffic parameters leading to lane-change related collisions," *Accident; Anal. Prevention*, vol. 38, no. 5, pp. 936–948, Sep. 2006.

[36] R. Yu and M. Abdel-Aty, "Utilizing support vector machine in real-time crash risk evaluation," *Accident; Anal. Prevention*, vol. 51, pp. 252–259, Mar. 2013.

[37] C. Xu, W. Wang, and P. Liu, "A genetic programming model for real-time crash prediction on freewaysays," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 574–586, Jun. 2013.

[38] M. Ahmed, M. Abdel-Aty, and R. Yu, "Assessment of the interaction between crash occurrence, mountainous freeway geometry, real-time weather and AVI traffic data," in *TRB Annu. Meet.*, Jul. 2011, p. 2450.

[39] C. Xu, P. Liu, W. Wang, and X. Jiang, "Development of a crash risk index to identify real time crash risks on freeways," *KSCE J. Civil Eng.*, vol. 17, no. 7, pp. 1788–1797, Oct. 2013.

[40] A. Vogt and J. G. Bared, "Accident models for two-lane rural roads: Segment and intersections," Federal Highway Admin., Washington, DC, USA, Tech. Rep. FHWA-RD-98-133, 2008.

[41] D. J. Sargent, "Comparison of artificial neural networks with other statistical approaches," *Cancer*, vol. 91, no. S8, pp. 1636–1642, 2001.

[42] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," *J. Biomed. Informat.*, vol. 35, nos. 5–6, pp. 352–359, 2002.

[43] A. Verikas, A. Gelzinis, and M. Bacauskiene, “Mining data with random forests: A survey and results of new tests,” *Pattern Recognit.*, vol. 44, no. 2, pp. 330–349, 2011.

[44] L. Rokach, “Ensemble-based classifiers,” *Artif. Intell. Rev.*, vol. 33, nos. 1–2, pp. 1–39, 2010.

[45] X. Li, D. Lord, Y. Zhang, and Y. Xie, “Predicting motor vehicle crashes using Support Vector Machine models,” *Accident Anal. Prevention*, vol. 40, no. 4, pp. 1611–1618, 2008.

[46] W. Wang, X. Qu, W. Wang, and P. Liu, “Real-time freeway sideswipe crash prediction by support vector machine,” *IET Intell., Transp. Syst.*, vol. 7, no. 4, pp. 445–453, 2013.

[47] N. Dong, H. Huang, and L. Zheng, “Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects,” *Accident Anal. Prevention*, vol. 82, pp. 192–198, Sep. 2015.

[48] H. Xiaoyu, W. Yisheng, and H. Siyu, “Short-term traffic flow forecasting based on two-tier K-nearest neighbor algorithm,” *Proc. Social Behav. Sci.*, vol. 96, pp. 2529–2536, Nov. 2013.

[49] L. Lin, Q. Wang, and A. W. Sadek, “A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction,” *Transp. Res. C, Emerg. Technol.*, vol. 55, pp. 444–459, Jun. 2015.

[50] D. Hand, D. Hand, H. Mannila, H. Mannila, P. Smyth, and P. Smyth, *Principles of Data Mining*, vol. 30. Cambridge, MA, USA: MIT Press, 2001.

[51] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “KNN model-based approach in classification,” in *Proc. Int. Conf. Cooperat. Inf. Syst. (CoopIS)*, 2003, pp. 986–996.

[52] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[53] T. K. Ho, “The random subspace method for constructing decision forests,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[54] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[55] Department for Transport. (2012). *GB Road Traffic Counts—Datasets—DGU*. [Online]. Available: http://data.gov.uk/dataset/gb-road-traffic-counts

[56] OpenStreetMap. (2016). *OpenStreetMap*. [Online]. Available: https://www.openstreetmap.org

[57] R. Dowling, A. Skabardonis, and V. Alexiadis, “Traffic analysis toolbox: Guidelines for applying traffic microsimulation modeling software,” Federal Highway Admin., Washington, DC, USA, Tech. Rep. FHWA-HRT-04-040, Jul. 2004, p. 146, vol. 3.

[58] *Traffic Modelling Guidelines TFL Traffic Manager and Network Performance Best Practice Version 3.0*, Transp. London, London, U.K., 2010.

[59] Wisconsin Department of Transportation. (2014). *Model Calibration—Traffic Analysis and Microsimulation*. [Online]. Available: http://www.wisdot.info/microsimulation/index.php?title=Model_Calibration

[60] F. Cunto, “Assessing safety performance of transportation systems using microscopic simulation,” Ph.D. dissertation, Dept. Civil Environ. Eng., Univ. Waterloo, Waterloo, ON, Canada, 2008.

[61] *MATLAB*, MathWorks, Natick, MA, USA, 2016.

[62] T. J. Triggs and W. G. Harris, “Reaction time of drivers to road stimuli,” *Med. Pregled*, vol. 62, pp. 9–114, Jun. 1982.

[63] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software,” *ACM SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[64] A. B. Hassanat, M. A. Abbadi, and A. A. Alhasanat, “Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 12, no. 8, pp. 33–39, 2014.

Christos Katrakazas (M’14) received the Diploma in civil engineering (transportation engineering cycle) from National Technical University of Athens, Greece, in 2013. He is currently pursuing the Ph.D. degree in intelligent vehicles with the School of Civil and Building Engineering, Loughborough University, U.K.

He has authored or co-authored one paper in an international journal and four international conference papers. His current research interests include autonomous vehicles, machine learning, and data mining applications for improved road safety.

Mohammed Quddus received the bachelor’s degree in civil engineering from Bangladesh University of Engineering and Technology in 1998, the master’s degree in transportation engineering from National University of Singapore in 2001, and the Ph.D. degree in intelligent transportation systems from Imperial College London in 2006

In 2006, he joined as a Lecturer with the School of Civil and Building Engineering, Loughborough University, U.K., and was promoted to a Senior Lecturer in 2010 and a Professor of intelligent transportation systems in 2013. He has authored over 100 technical papers in international refereed journals and conference proceedings. His current research interests include high-accuracy and integrity land vehicle navigation, autonomous navigation, and sensor fusion.

Prof. Quddus is a member of the U.S. Transportation Research Board, the British Computer Society, the EPSRC, and the Universities’ Transport Study Group, U.K.

Wen-Hua Chen (M’00–SM’06) received the M.Sc. and Ph.D. degrees from Northeast University, Shenyang, China, in 1989 and 1991, respectively. From 1991 to 1996, he was with the Department of Automatic Control, Nanjing University of Aeronautics and Astronautics, China. From 1997 to 2000, he was a Researcher and then a Lecturer of control engineering with the Centre for Systems and Control, University of Glasgow, U.K. In 2000, he was a Lecturer with the Aeronautical and Automotive Engineering Department, Loughborough University, U.K., and was appointed as a Professor in 2012.

His research interests include the development of advanced control, signal processing and decision-making methods, and their applications for unmanned vehicles.

Prof. Chen is a Chartered Engineer in U.K., a fellow of the Institution of Engineering and Technology, and a Fellow of the Institution of Mechanical Engineers.