TIME SERIES CLASSIFICATION USING IMBALANCED LEARNING FOR REAL-TIME SAFETY ASSESSMENT

Christos Katrakazas, PhD
Post-Doctoral Research Associate
Technical University of Munich
Arcisstrasse 21, 80333, Munich, Germany
Tel: +49(0)8928910463
Email: c.katrakazas@tum.de

Constantinos Antoniou, PhD
Professor
Technical University of Munich
Arcisstrasse 21, 80333, Munich, Germany
Tel: +49(0)8928910460
Email: c.antoniou@tum.de

George Yannis, PhD
Professor
National Technical University of Athens
Department of Transportation Planning and Engineering
5 Iroon Polytechniou St., GR-15773, Athens, Greece
Tel: +302107721326

Word count: 7,231 words text + 1 table x 250 words = 7,481 words

ABSTRACT

The probability of estimating a traffic collision happening in real-time primarily depends on comparing traffic conditions just before a collision with traffic conditions during normal operations. Most studies however utilize aggregated traffic data and are not concerned with the dynamic nature of collisions or the imbalance of safety databases which can lead to erroneous real-time predictions. In this study, this is overcome through the use of raw speed time series data of variant duration (i.e. 1-minute to 5-minute time series data) from a driving simulator experiment and the use of imbalanced learning techniques. Two classifiers are then employed to examine the proposed idea: (i) Random Forests (RFs) – an ensemble classifier and (ii) Neural Networks (NNs) – a popular classifier in the literature. These classifiers are tested on the original time series data, as well as on time-series treated with the imbalanced learning techniques of undersampling and its integration with oversampling. The main results demonstrate the viability of using raw speed time series data for real-time safety assessment and the superiority of time series with 4-minute duration in the classification results. Furthermore, RFs perform well even on 1-minute time series data while the classification results can be enhanced by up to 40% from imbalanced learning approaches. It is also demonstrated that the classification results outperform similar approaches in the literature. However, real-world traffic data and the use of more sophisticated classifiers (e.g. Deep Learning) are expected to provide more effective collision predictions.

Keywords: Real-time Collision Prediction, Time-series Classification, Imbalanced Learning, Neural Networks, Random Forests

INTRODUCTION

Real-time collision prediction is formulated on the basis that the probability of a collision occurring could be estimated for a short-time prediction horizon from traffic data retrieved online (*1*). Intensive research has taken place in the past two decades to make real-time collision prediction more accurate regardless of the traffic data used for the analysis. However, current models need further enhancing. Traditional real-time collision prediction models usually follow four steps: i) select actual traffic variables (e.g. temporal or spatial means and variance of them) as predictors, ii) collect data corresponding to historical collision cases and normal traffic conditions, iii) formulate a classification problem and utilize a collision prediction model to estimate the probability of a collision and iv) evaluate the modelling performance. Nevertheless, efficiently applying these four steps is not perfectly tractable and the dynamic evolution of the accidents is not taken into account. Traffic data might not be available at all times and hence classifiers need to be able to work with limited or bad quality data (*2*). Machine learning classifiers have been applied to solve the problem of correlated variables and missing data, however, in most cases they act like "black-boxes" which restrict the interpretability of the models. Moreover, as collisions are rare events, the data collection of collision-prone and normal traffic cases leads to an overrepresentation of cases the cases representing normal traffic which, consequently, results in biased classifiers and a large number of false alarms (*3*). Classification of imbalanced datasets is a documented problem in data mining (*4–6*). The most important problem with imbalanced data is the high misclassification rate for the under-represented class, because the classifier favours the majority class. Oddly, there is little evidence in the literature to date to take into account the dynamic nature of accidents as well as the imbalance of collision datasets when building real-time collision prediction classifiers.

This paper will therefore attempt to classify time series speed data and simultaneously assess the potential enhancement in real-time collision prediction models after treating datasets with imbalanced learning techniques. Two machine learning classifiers, Neural Networks(NNs) and Random Forests (RFs) are used for classifying speed time series as collision-prone or safe using different temporal intervals.

The paper is organized as follows: firstly, the existing literature and its main findings are reviewed. A detailed description of the RFs and NNs classification algorithms is described next, along with principles of imbalanced learning. This is followed by a presentation of the data used in the analysis, the pre-processing methodology and the results of the classification algorithm. Finally, the last section summarizes the main conclusions of the study and gives recommendations for future research.

LITERATURE REVIEW

Most recent approaches in real-time collision prediction modelling require the utilization of data just before a collision occurrence (termed as collision-prone) as well as data of collision-free (also termed as normal) traffic conditions. Traffic data resembling collision-prone and normal traffic are usually employed as a matched-case control methodology, in which every collision-prone traffic condition is matched with a number of normal traffic cases. This is so as to single out collision precursors (i.e. traffic indications of an imminent collision). The technique of matched-case control for real-time collision prediction studies was initially introduced by Abdel-Aty et al. (*7*) and has thereafter been used massively because it eliminates the effects of location, time and weather conditions on the probability of a collision occurrence. In studies employing matched-case control research design, the ratio of collision-prone to safe traffic conditions varies from 1:4 (*8*) and 1:5 (*9*, *10*) to 1:34 (*11*). In the literature, there is no set rule for choosing a ratio between cases and controls as normally the number depends on the available data. However, according to (*12*) ratios greater than 1:5 do not result in a statistically significant difference in predicting performance.

Temporal Aggregation of traffic data and dynamic considerations

As the application of real-time collision prediction models is the proactive identification of collision-prone traffic conditions, researchers aggregate the raw traffic data coming from various traffic sensors into different intervals of temporal aggregations. In (*13*), for instance, aggregated traffic data into 5-minute intervals and suggested that 5 minutes just before the collision occurrence should represent *hazardous* traffic conditions while 30 minutes of aggregated traffic data before the crash should imply *safe* traffic. 2.5 minutes of data just before the collision event were discarded in (*14*) 30 minutes of aggregated traffic data for modelling real-time collision risk were utilized. Abdel-Aty et al. (*1*) stated that raw data (e.g. 20-second, 30-second or 1-minute data) from loop detectors or other traffic measuring devices include random noise and therefore their utilization in collision prediction modelling is burdensome. They divided the 30-minute interval just before a collision into six 5-minute time intervals and concluded that the best results for collision prediction are obtained using traffic data 5-10 minutes before a collision. The same authors (*15*) utilized 3-minute traffic data aggregation and concluded that it performed worse than 5-minute aggregation.

In the following years, the vast majority of the literature on real-time collision prediction (*11*, *16–25*) followed similar methodologies; traffic data are aggregated in 5-minute intervals and the five-minute interval 5-10 minutes before the crash is used for predicting if a collision is imminent or not. The only differentiations from the majority of studies were found in (*26*) who utilized traffic data from the interval 0-5 minutes before the collision and (*27*, *28*), where traffic data from the interval 10-15 minutes before each collision were used for modelling. The prediction of each approach is relative to the traffic data used to calibrate the model. For example, if the model is calibrated using data 5-10 minutes before the collision, the model would be able to identify whether the traffic conditions at a specific time moment are hazardous enough to cause a collision in the next 10 minutes.

More recently, (*29*) attempted to correlate collision risk with microscopic traffic data (raw loop detector data) along with surrogate safety measurements (e.g. Time-to-Collision or TTC). However, their focus was the identification of weather and kinematic characteristics leading to fog-related collisions only and not the identification of collision-prone traffic conditions. A large dataset of highly disaggregated AVI data from two motorways in Chile was used in (*30*), but also aggregated their dataset into 5-minute intervals to cope with the influence of geometric or driving behaviour characteristics on the prediction performance. Recently, Katrakazas (*31*, *32*) utilized highly disaggregated data for real-time safety assessment, however the majority of the disaggregated data were simulated.

Methods utilized for analysis

Methodologically, recent real-time collision prediction approaches are divided into two broad categories: (1) statistical (*34*, *35*) and (2) artificial intelligence (AI) or machine learning (*11*, *36–40*).

With regards to statistical approaches, traditional binary logit (*1*) and Bayesian logit; (*18*) as well as random parameters logit models (*2*) have been applied. In a traditional logit model (i.e. with fixed effects) the estimated coefficients correspond to averaged effects without considering individual diversity. Random parameter models can account for the heterogeneity of road geometry, weather conditions or driving behaviour and have superior performance when compared to traditional logit (*41*).However, regression models require the determination of a critical odds ratio as a threshold for the identification of collision-prone traffic conditions (*42*) 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, a number of studies (*43*, *44*) utilized three types of NNs: (i) Probabilistic (*36*), (ii) Radial Basis Function (*44*, *43*) and (iii) Multilayer Perceptron (*44*, *43*) for real-time collision estimation on American freeways, demonstrating that NNs which do not require any distributional assumptions outperform statistical approaches. NNs usually require a large dataset for training (*45*). However, their major drawback is related to the incorporation of the “black-box” effect, which prevents clear understanding of the model’s underpinning properties, interpretation of the model’s results and model transferability (*46*). Furthermore, NN models often suffer from over-fitting (*38*) and require extra computational resources to overcome it (*45*). The same “black-box” effect was also documented as a problem for other machine-learning approaches such as Support Vector Machines (SVMs) (*38*).

Genetic Programming, an extension of Genetic Algorithms (*47*), was proposed by Xu et al. (*27*) to remove the “black-box” effect of machine learning approaches, but their model faced difficulties with regards to transferability and practical implementation. In another attempt to tackle the effect of “black-box”, Lv et al. (*48*) and Lin et al. (*28*) utilized the non-parametric algorithm of k-Nearest Neighbours (k-NN).

In order to deal with the drawbacks of previous approaches (both logistic regression and machine learning ones), Hossain and Muromachi proposed Bayesian Networks (*11*). They investigated collision prediction on main motorway segments and ramp vicinities by using traffic flow variables and finding an ideal arrangement of detectors for data collection, after hypothesizing that the collision mechanism is different on main segments and ramps. Their study, however, had limited transferability.

Sun and Sun (*40*) implemented Dynamic Bayesian Networks, an extension of Bayesian Network able to model temporally sequential data. The focal point of their approach is that they treated collisions as an event triggered by dynamically changing precursors, which is a more realistic view of investigating collision probability over focusing on making point predictions based on aggregated traffic data. Bayesian Networks combine the probability and the graph theory to represent dependencies between predictors and the dependent variable. In order to be able to represent the probabilities of each of the included variables, Bayesian Networks require a sufficiently large dataset which makes them difficult to be implemented with small and unbalanced datasets. On the same principle, Theofilatos (*49*, *50*) indicated that time series could be applied for real-time safety applications due to the dynamic nature of collisions and utilized SVMs after applying discrete wave transformation to 3-hour long (aggregated at 5-minute intervals) time series data before accident occurrences. More recently, Fountas et al. (*51*) developed dynamic random parameters models, but also utilized aggregated traffic data.

Finally, only the work of (*30*) took into consideration the imbalance of safety related datasets and applied a technique of imbalanced learning with SVMs to predict the probability of a collision in real-time. However, as mentioned before the data used were aggregated and there was no comparison with other machine learning approaches.

It can be observed from the literature review, that the state-of-the-art in real-time safety assessment, fails to utilize disaggregated data and also lacks in treating the imbalance of safety datasets and the dynamic nature of collision occurrences. Therefore, the current paper will attempt to classify time series data, of variant duration and by making use of imbalanced learning approaches.

METHODOLOGY

Binary classification and its evaluation metrics in real-time collision studies

The main objective of this study is to identify collision-prone speed time series from highly disaggregated data by using the RF and NN classifiers. As this objective aims to distinguish between two classes (i.e. collision-prone and safe speed), the problem is a binary classification one.

Consider a training dataset $X_{training} = \{(x_n, y_n), n = 1, \dots N\}$ being available where x_n is a predictor variable and y_n={0,1} is a response. A binary classification problem is the one attempting to build a function *f* which, given new data instances will assign them to the correct class. Moreover, the classification performance of every classifier is initially assessed through the confusion matrix. In a confusion matrix, the predictions of each data instance are contrasted with the original class to which they belonged, so as to ascertain whether they are correctly classified. In the real-time collision prediction task, the binary classification problem is concerned with the identification of collision-prone traffic, hence the positive class represents "*collision-prone*" traffic and the negative class represents "*safe"* traffic.

Usually, classification performance is measured with the confusion matrix which demonstrates the quantities of correctly identified and misclassified instances for each of the two classes.

Based on the confusion matrix, widely used metrics include:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (1)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

$$\text{G-means}= \sqrt{Recall * Specificity} \quad (4)$$

$$\text{f1-measure}= \frac{2*precision*recall}{precision+recall} \quad (5)$$

where: TN: True Negative, TP: True Positive, FN: False Negative, FP: False Positive.

The recall statistic shows the correct classification accuracy with respect to collision-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 the classes. G-means is used to ensure whether the use of an imbalance dataset has any negative impact on the balanced qualification accuracy. Lastly, the *f1*-measure is a metric which resembles the collision-prone classification ability of the classifier models (*40*).

Classification algorithms

RFs belong 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 (*55*). 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 (*55*).

A NN is a parallel-distributed processor made up of simple processing units having natural propensity for learning from an available training dataset and making general predictions for future “unknown” data (*56*). This generalization property of NNs refers to the ability of a “trained” network to provide satisfactory responses even for inputs that were not used for training. In order to define a NN models three entities need to be defined: i) the model of processing elements themselves, ii) the network topology, and the learning rules. In this study, a multi-layer perceptron (MLP) network with feed-forward connections was used.

Imbalanced learning

One of the primary limitations of real-time collision prediction models as indicated in the literature review is the imbalance of the datasets used in real-time collision prediction modelling where safe traffic condition cases are over-illustrated against collision-prone conditions due to the rarity of collision events. This subsection will discuss the methods used to improve the performance of real-time collision prediction classifiers.

Classification of imbalanced datasets is a documented problem in data mining (*4–6*). The most important problem with imbalanced data is the high misclassification rate for the under-represented class, because the classifier favours the majority class. To overcome this problem proposed solutions from the literature can be grouped into three groups:

1) Data sampling
2) Algorithm alteration
3) Cost-sensitive learning

The first solution requires that the sampling of training cases should be modified to a certain extent, in order for a more balanced dataset to be produced. Next, the algorithm alterations solution relates to modifications made in learning algorithms e.g. in the kernels for kernel-based approaches such as SVMs or in the construction of trees for tree-based approaches such as Random Trees or RFs. The third solution applies higher misclassification costs for instances of the minority class (i.e. for false positives) and lower misclassification costs for the majority class (i.e. for false negatives). In this study the first solution will be utilized (i.e. Data Sampling) and is described in the following sub-section.

Data Sampling

In order to achieve a more balanced dataset, He and Garcia (*4*) propose random oversampling or undersampling. Random oversampling is a technique which artificially appends data in the original dataset while random undersampling is a technique that randomly selects cases from the majority class so that a more balanced dataset is acquired. However, it is suggested in (*4*)that oversampling might lead to over-fitting. Thus, undersampling would generally be preferable for the purposes of this work. However, data cleansing in conjunction with oversampling is also suggested as a solution to address over-fitting and hence it will also be tested.

Reviewing the literature in undersampling and overasampling with data cleansing, it was found that Repeated Edited Nearest Neighbours (RENN) (*57*), its integration with Synthetic Minority Oversampling TEchnique (SMOTE) (*58*) performed well for classes that are difficult to recognise (*59*).

RENN utilizes the Edited Nearest Neighbour (ENN) algorithm (*60*) repeatedly until all the instances in the dataset have a majority of their neighbours within the same class. ENN applies the kNN algorithm and removes all misclassified instances from the training dataset. In this way, the difference between classes is more obvious and a smooth decision threshold is obtained. The RENN algorithm developed in (*61*)is briefly discussed below:

- If D_e is the dataset acquired from the ENN algorithm and D_o is the original dataset repeat:
    - At every iteration i for each instance x_i in D_e discard x_i if it is misclassified using kNN
- Until ${D_e}^i = {D_e}^{i-1}$ where ${D_e}^i$ is the edited dataset in iteration i and ${D_e}^{i-1}$ is the edited dataset in Iteration i-1.

SMOTE integrated with ENN aims at producing well-defined class clusters which can potentially improve classification results. After artificially generating instances of the minority class through SMOTE, ENN is implemented to conduct the data cleaning in depth and removes data instances from both classes when the three nearest neighbours of a data instance are misclassified (*59*). This is beneficial, especially for datasets with a small number of instances in the positive class, for instance collision-prone traffic, in datasets containing collision data which are rare events. The algorithm will be henceforth termed as SMOTE-ENN.

DATA DESCRIPTION AND PRE-PROCESSING

The data utilized in this study were collected using a driving simulator at the Department of Transportation Planning and Engineering of the National Technical University of Athens. More specifically, a FOERST Driving Simulator FP, consisting of 3 LCD wide screens 40" (Full HD: 1920x1080 pixels, a driving position and a support motion base was employed. The simulator's dimensions at full development are 230x180cm, the width of its base is 78cm and its total field of view is 170 degrees. The data collected with the simulator were originally used for the Distract project (*62*) which investigated the causes and impacts of driver distraction, using a driving simulator.

The driving scenarios included driving in rural, urban and motorway environments. For the purposes of this paper only the rural area data were used. Each experiment included a 15- to 20-minute warm-up drive, so as to familiarize the driver with the simulator, and a 20-minute recorded driving session. The rural route was 2.1 km long on a single carriageway, with 3m lane width, zero gradient and mild horizontal curves. During each trial, 2 unexpected incidents were programmed to occur and concerned the sudden appearance of an animal. Only incidents resulting in crashes were considered in this study. The experiment was counterbalanced with regards to the number and order of trials. For more details on the dataset and the experiment, the reader is referred to (*62*, *63*).

In total, 279 driving sessions were taken into account for the current paper. For every driving session the variables of interest were the actual vehicle speed in km/h, and the binary existence of a crash or not (1 for crash, 0 for safe driving). Measurements were recorded every 17 and 33 milliseconds. The sessions were divided into those that included a collision event and those that did not. In order to obtain the time series for collision events, the collision time was initially identified, and speed measurements were taken into account for 1,2,3,4 and 5 minutes before each collision, and were labelled as "collision-prone". The five different time intervals were chosen so as to investigate the effect of time series length on the classification results. In order to represent safe driving conditions, the sessions that did not included collisions were marked as safe and were divided into 1- to 5-minute time series. 169 collision events were found in the dataset, and the ratio of collision: safe time series was 1:2, 1:3, 1:4, 1:6 and 1:13 for the 5-,4-,3-,2- and 1-minute time series respectively.

RESULTS AND DISCUSSION

As mentioned previously, the algorithms tested were RFs and NNs. Before the initiation of each algorithm, an optimization routine was run along with 10-fold cross-validation in order to find the optimal parameters for each algorithm using the training dataset. RFs with 100 estimators of maximum depth 3 and a Multilayer Perceptron with α=0.05 were utilized. In order to avoid over-fitting and assure optimal results, 2/3 of the dataset were used for training the classifiers and 1/3 of the dataset was used for testing the classification results. The models were developed in Python 2.7 using the scikit-learn (*64*) package. The *imbalanced-learn* package in python (*65*) was utilized in order to apply imbalanced learning approaches to the dataset. The two techniques that were utilized were RENN regarding undersampling and the combination of SMOTE (taking into account 10 data neighbours) and ENN. Each classification algorithm (i.e. RFs and NNs) was trained with the balanced dataset and its performance was tested on the original (imbalanced) dataset. By testing the performance on the original dataset, it is ensured that the validation of the classification results is not based on artificially created instances from SMOTE-ENN or a smaller sample acquired through RENN, but is directly acquired from the original dataset.

The classification results evaluated through equations 1-5, are presented in Table 1. For every algorithm, a number is used to denote the length of the time series (e.g. 1 denotes a time series of 1-minute duration) and in the cases where imbalanced learning has been used the technique implemented is also indicated. For example, RF_3_RENN denotes the classification results for the Random Forest Classifier on 3-minute time-series data which have been treated with the imbalanced technique of RENN to counteract on the imbalance between collision-prone and safe instances.

TABLE 1 Classification metrics for the developed classifiers

Classifier	Accuracy	Precision	Recall	Specificity	f1-score	G-Means	False Alarm Rate
RF_1	95.13%	70.00%	43.75%	98.70%	53.85%	55.34%	1.30%
NN_1	93.50%	-	0.00%	100.00%	-	-	0.00%
RF_2	91.25%	88.24%	50.85%	98.74%	64.52%	66.98%	1.26%
NN_2	68.97%	30.14%	74.58%	67.92%	42.93%	47.41%	32.08%
RF_3	92.49%	87.50%	65.12%	98.10%	74.67%	75.48%	1.90%
NN_3	84.98%	54.24%	74.42%	87.14%	62.75%	63.53%	12.86%
RF_4	91.67%	88.89%	71.11%	97.48%	79.01%	79.50%	2.52%
NN_4	87.25%	85.19%	51.11%	97.48%	63.89%	65.98%	2.52%
RF_5	83.13%	86.11%	58.49%	95.33%	69.66%	70.97%	4.67%
NN_5	66.88%	0.00%	0.00%	100.00%	-	-	0.00%
RF_1_RENN	96.02%	75.91%	61.54%	98.56%	67.97%	68.35%	1.44%
NN_1_RENN	93.38%	60.71%	10.06%	99.52%	17.26%	24.71%	0.48%
RF_2_RENN	92.82%	75.48%	69.23%	96.50%	72.22%	72.29%	3.50%
NN_2_RENN	83.65%	23.53%	9.47%	95.21%	13.50%	14.93%	4.79%
RF_3_RENN	92.63%	95.73%	66.27%	99.26%	78.32%	79.65%	0.74%
NN_3_RENN	79.90%	50.00%	0.59%	99.85%	1.17%	5.44%	0.15%
RF_4_RENN	91.30%	89.29%	73.96%	97.05%	80.91%	81.26%	2.95%
NN_4_RENN	52.36%	33.41%	91.72%	39.29%	48.97%	55.35%	60.71%
RF_5_RENN	85.71%	76.27%	79.88%	88.43%	78.03%	78.06%	11.57%
NN_5_RENN	68.23%	-	0.00%	100.00%	-	-	0.00%
RF_1_SMOTE-ENN	88.75%	36.22%	84.02%	89.10%	50.62%	55.17%	10.90%
NN_1_SMOTE-ENN	78.40%	21.24%	79.29%	78.33%	33.50%	41.03%	21.67%

RF_2_SMOTE-ENN	89.23%	56.91%	82.84%	90.23%	67.47%	68.66%	9.77%
NN_2_SMOTE-ENN	58.69%	23.84%	94.08%	53.18%	38.04%	47.36%	46.82%
RF_3_SMOTE-ENN	92.87%	80.45%	85.21%	94.79%	82.76%	82.79%	5.21%
NN_3_SMOTE-ENN	29.25%	21.68%	96.45%	12.35%	35.40%	45.72%	87.65%
RF_4_SMOTE-ENN	92.48%	80.10%	92.90%	92.34%	86.03%	86.26%	7.66%
NN_4_SMOTE-ENN	89.97%	73.27%	94.08%	88.61%	82.38%	83.03%	11.39%
RF_5_SMOTE-ENN	83.65%	69.16%	87.57%	81.82%	77.28%	77.82%	18.18%
NN_5_SMOTE-ENN	76.13%	61.05%	68.64%	79.61%	64.62%	64.73%	20.39%

From Table 1 it can be observed that imbalanced learning significantly enhances the performance of the classifiers. RFs generally outperform NNs, while the best results are indicated for the 3- and 4-minute time series data. When comparing the imbalanced learning techniques, it is observed that the integration of oversampling with undersampling results in better classification performance, than only undersampling the majority class. Observing the original time series data, without any imbalanced learning treatment, it is shown that NNs perform better in terms of identifying correctly, collision-prone conditions in shorter time series (i.e. consisting of 1-minute, 2-minute and 3-minute measurements) while RFs perform better when the duration of time series increases. The majority of false alarms is higher for NNs than RFs, which can be explained by the small data size, as usually NNs fail to perform well when small datasets are employed (*66*). Looking at the overall performance of the classifiers, as depicted in the f1-score and G-means, it is understood that the best results are obtained with RFs and SMOTE-ENN using a 4-minute time series (f1=86%, G-means=86.3%) and guaranteeing correct identification of both collision-prone speed time series, as well as safe conditions. More importantly, it is also shown that even without imbalanced treatment, the original 4-minute series outperforms the majority of the developed classifiers. Another important finding is that RFs when combined with undersampling are able to identify almost 70% of collision-prone speed conditions with a very small false alarm rate even when 1-minute or 2-minute speed data are used. As a result, even when data collected over a short time period are available, they can efficiently be used for real-time safety assessment, thus improving the speed of predictions.

To further illustrate the enhancement in classification performance, that imbalance learning provides, Figure 1 demonstrates the percentage change in recall (i.e. the identification of collision-prone speed conditions), f1and G-means scores for the RF classifier over all the speed time series used.

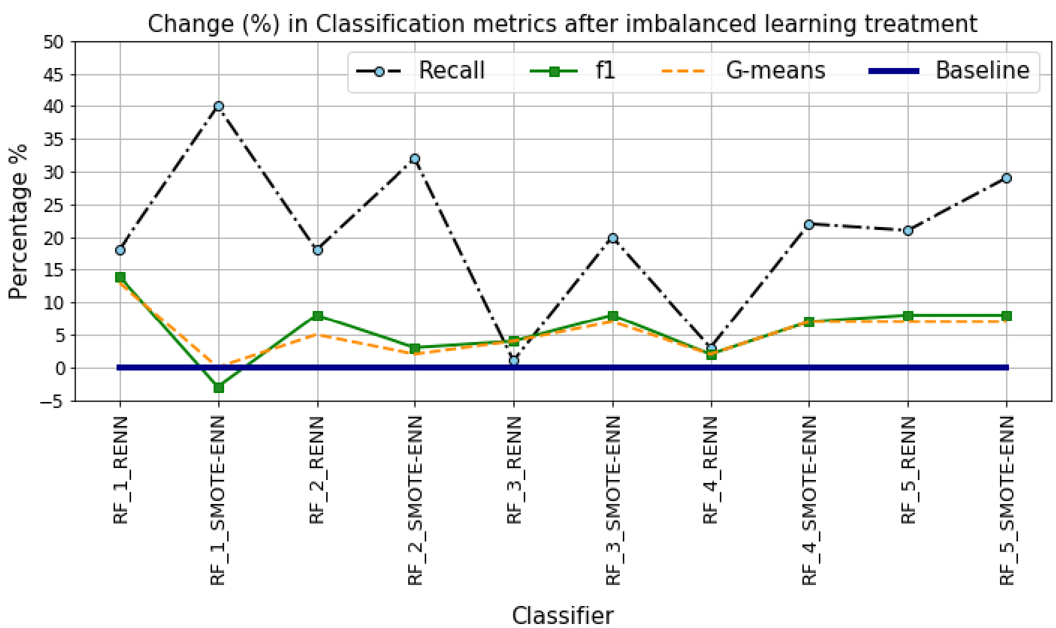

FIGURE 1: Percentage change in classification metrics when compared with the untreated classification results

Observing Figure 1, it is shown that the improvement in identifying collision-prone conditions, is higher for time series of shorter duration, and can assist in recognizing 40% more "dangerous" cases. Such an enhancement is also shown when 4-minute or 5-minute series are used, however the effect is only half when compared to 1-minute or 2-minute speed time series. Nevertheless, it is also demonstrated that although the enhancement in recall is significant, the overall classification performance is only marginally improved after treating with imbalanced learning techniques.

In order to correlate the current paper with existing literature, the best of the developed speed time series classifiers are compared to the results of Theofilatos et al. (*49*, *50*), who also investigated speed-time series classification for estimating accident involvement, but using 5-minute aggregated data in 3-hour long time series. The comparison (illustrated on Figure 2) demonstrates that the time-series classifiers developed in this work, outperform the ones already developed in the literature in identifying collision-prone speed conditions and with regards to false alarm rates. Even classifiers without the treatment of imbalanced learning utilizing raw 4-minute time series data perform similarly with the classifiers developed in (*49*, *50*), which utilized real-world data over a period of 3-hours.

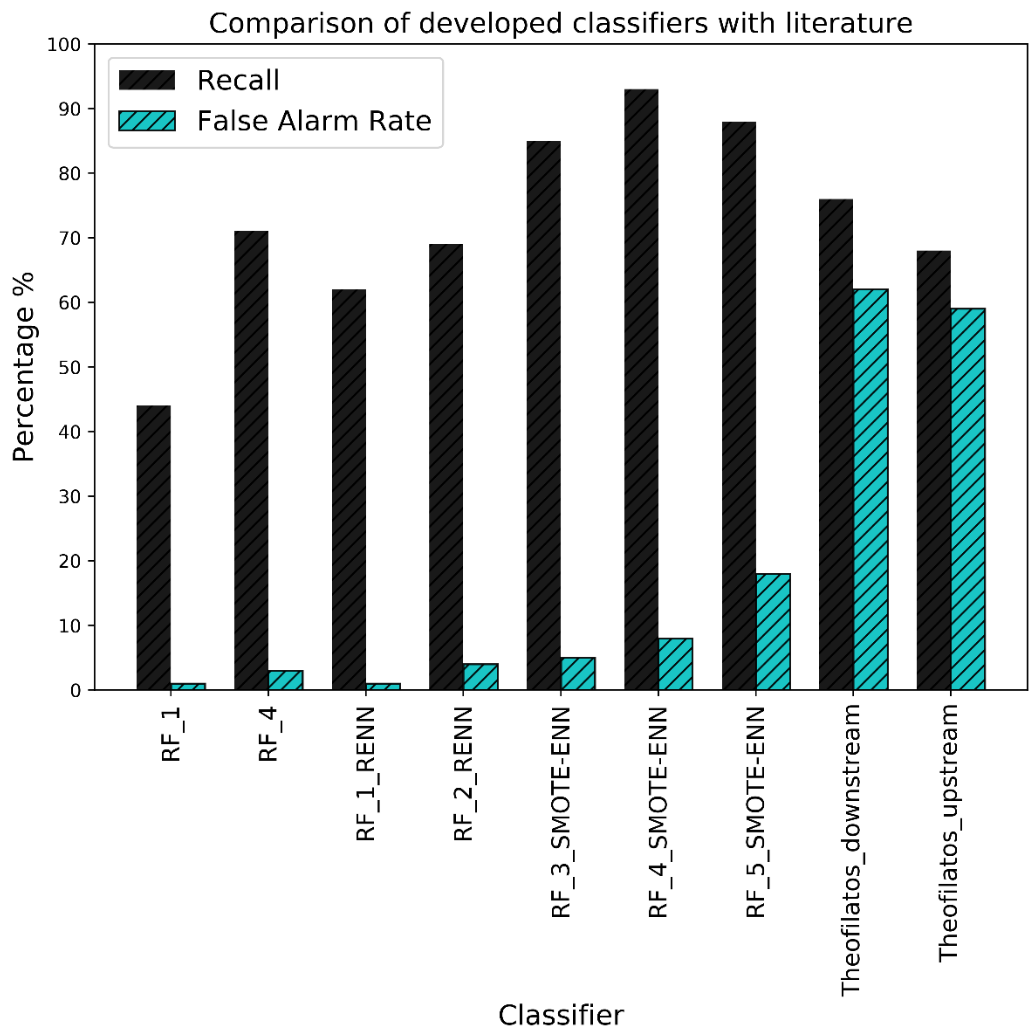

FIGURE 2 Comparison of developed classifiers with the results of Theofilatos et. al (2018)

CONCLUSIONS

Real-time collision prediction has been the aim of many experts in the area of ITS over the recent decades. However, most of the approaches fail to take into account the dynamic nature of collision occurrences, rely on aggregated data which fail to efficiently reflect the specific traffic dynamics that may lead to collisions and do not consider the imbalance of safety databases.

This paper proposes the classification of raw speed time series data before collision events using imbalanced learning. The approach intents to overcome two problems: i) the use of aggregated data, as raw time-series data are utilized and ii) the imbalance of safety databases through the use of imbalanced learning. Two classification algorithms, the renowned NNs and the ensemble RFs were utilised to distinguish between collision-prone and safe speed conditions. The data used were obtained through a driving simulator experiment in Athens, which took place in order to assess the driving performance of drivers with cognitive impairment. Five speed time series were constructed in order to test the effect of duration in the classification results. Regarding imbalanced learning, two techniques were utilized, namely undersampling of the majority class (i.e. safe traffic conditions) and oversampling of the minority class (i.e. collision-prone traffic) integrated with undersampling. The imbalanced learning classifiers were trained using balanced datasets and were tested on the original imbalanced datasets. The algorithms' performance was evaluated using their overall accuracy and the metrics of recall, specificity, precision, recall, G-means and F-measure.

The classification results showed that raw speed time series can efficiently be used in real-time safety assessment. The classification performance of the developed classifiers outperforms results in the literature in terms of identifying collision-prone speed conditions with a low false-alarm rate. It was shown that RFs in general lead to better classification results when compared to NNs, and the treatment with imbalanced learning can enhance results up to 40% even when 1-minute time series are utilized for real-time classifications.

However, in order for the proposed approach to become more efficient tests should be performed with real-world data in order to obtain traffic conditions as much realistic as possible. Lastly, more sophisticated techniques such as Bayesian Networks or Deep Learning should be explored to cope with the noise of time series of shorter duration.

ACKNOWLEDGEMENT

The research was funded by the EU H2020 NOESIS Project (Project Number: 769980)

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: Study conception and design: C. Katrakazas, data collection: C. Antoniou, G. Yannis; analysis and interpretation of results: C. Katrakazas; draft manuscript preparation: C. Katrakazas, C. Antoniou, G. Yannis. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Abdel-Aty, M., and A. Pande. The Viability of Real-Time Prediction and Prevention of Traffic Accidents. In *Efficient Transportation and Pavement Systems* (Al-Qadi, T. Sayed, Alnuaimi, and Massad, eds.), Taylor & Francis Group, London, pp. 215–226.
2. Xu, C., W. Wang, P. Liu, and Z. Li. Calibration of Crash Risk Models on Freeways with Limited Real-Time Traffic Data Using Bayesian Meta-Analysis and Bayesian Inference Approach. *Accident Analysis and Prevention*, Vol. 85, 2015, pp. 207–218. https://doi.org/10.1016/j.aap.2015.09.016.
3. Xu, C., P. Liu, and W. Wang. Evaluation of the Predictability of Real-Time Crash Risk Models. *Accident Analysis and Prevention*, Vol. 94, 2016, pp. 207–215. https://doi.org/10.1016/j.aap.2016.06.004.
4. He, H., and E. A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 9, 2009, pp. 1263–1284. https://doi.org/10.1109/TKDE.2008.239.
5. Sun, Y., A. K. C. Wong, and M. S. Kamel. Classification of Imbalanced Data : A Review. *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 23, No. 4, 2009, pp.

687–719. https://doi.org/10.1142/S0218001409007326.

6. López, V., A. Fernández, S. García, V. Palade, and F. Herrera. An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics. *Information Sciences*, Vol. 250, 2013, pp. 113–141. https://doi.org/10.1016/j.ins.2013.07.007.
7. Abdel-Aty, M., N. Uddin, A. Pande, F. Abdalla, and L. Hsia. Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression. *Transportation Research Record,Journal of the Transportation Research Board*, 2004.
8. Ahmed, M., and M. Abdel-Aty. A Data Fusion Framework for Real-Time Risk Assessment on Freeways. *Transportation Research Part C: Emerging Technologies*, Vol. 26, 2013, pp. 203–213. https://doi.org/10.1016/j.trc.2012.09.002.
9. Abdel-Aty, M., A. Pande, A. Das, and W. J. Knibbe. Assessing Safety on Dutch Freeways with Data from Infrastructure-Based Intelligent Transportation Systems. *Transportation Research Record*, Vol. 2083, No. 2083, 2008, pp. 153–161. https://doi.org/10.3141/2083-18.
10. Ahmed, M. M., and M. Abdel-Aty. The Viability of Using Automatic Vehicle Identification Data for Real-Time Crash Prediction. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 13, No. 2, 2012, pp. 459–468. https://doi.org/10.1109/TITS.2011.2171052.
11. Hossain, M., and Y. Muromachi. A Bayesian Network Based Framework for Real-Time Crash Prediction on the Basic Freeway Segments of Urban Expressways. *Accident; analysis and prevention*, Vol. 45, 2012, pp. 373–81. https://doi.org/10.1016/j.aap.2011.08.004.
12. Roshandel, S., Z. Zheng, and S. Washington. Impact of Real-Time Traffic Characteristics on Freeway Crash Occurrence: Systematic Review and Meta-Analysis. *Accident Analysis and Prevention*, Vol. 79, 2015, pp. 198–211. https://doi.org/10.1016/j.aap.2015.03.013.
13. Oh, C., J.-S. Oh, S. G. Ritchie, and M. Chang. *Real-Time Estimation of Freeway Accident Likelihood, Report UCI-ITS-TS-WP-00-8, Department of Civil and Environmental Engineering and Institute of Transportation Studies, University of California, Irvine*. 2001.
14. Golob, T. F., and W. W. Recker. A Method for Relating Type of Crash to Traffic Flow Characteristics on Urban Freeways. *Transportation Research Part A: Policy and Practice*, Vol. 38, No. 1, 2004, pp. 53–80. https://doi.org/10.1016/j.tra.2003.08.002.
15. Pande, A., and M. Abdel-Aty. A Freeway Safety Strategy for Advanced Proactive Traffic Management. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, Vol. 9, No. 3, 2005, pp. 145–158. https://doi.org/10.1080/15472450500183789.
16. Ahmed, M., M. Abdel-Aty, and R. Yu. Assessment of the Interaction between Crash Occurrence , Mountainous Freeway Geometry , Real-Time Weather and AVI Traffic Data. *TRB 2012 Annual Meeting*, Vol. 2450, No. July 2011, 2012.
17. Shew, C., A. Pande, and C. Nuworsoo. Transferability and Robustness of Real-Time Freeway Crash Risk Assessment. *Journal of safety research*, Vol. 46, 2013, pp. 83–90. https://doi.org/10.1016/j.jsr.2013.04.005.
18. Yu, R., M. Abdel-Aty, and M. Ahmed. Bayesian Random Effect Models Incorporating Real-Time Weather and Traffic Data to Investigate Mountainous Freeway Hazardous Factors. *Accident; analysis and prevention*, Vol. 50, 2013, pp. 371–6. https://doi.org/10.1016/j.aap.2012.05.011.
19. Hassan, H. M., and M. a Abdel-Aty. Predicting Reduced Visibility Related Crashes on Freeways Using Real-Time Traffic Flow Data. *Journal of safety research*, Vol. 45, 2013, pp. 29–36. https://doi.org/10.1016/j.jsr.2012.12.004.
20. Wu, Y., H. Nakamura, and M. Asano. A Crash Risk Estimation Model for Urban Expressway Basic Segments Considering Geometry , Traffic Flow and Ambient Conditions. *Eastern Asia Society for Transportation Studies*, Vol. 9, 2013.
21. Xu, C., 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, 2015, pp. 28–35. https://doi.org/10.1080/15389588.2014.909036.
22. Fang, S., W. Xie, J. Wang, and D. R. Ragland. Utilizing the Eigenvectors of Freeway Loop Data Spatiotemporal Schematic for Real Time Crash Prediction. *Accident Analysis and Prevention*, Vol. 94, 2016, pp. 59–64. https://doi.org/10.1016/j.aap.2016.05.013.

23. Xu, C., P. Liu, B. Yang, and W. Wang. Real-Time Estimation of Secondary Crash Likelihood on Freeways Using High-Resolution Loop Detector Data. *Transportation Research Part C: Emerging Technologies*, Vol. 71, 2016, pp. 406–418. https://doi.org/10.1016/j.trc.2016.08.015.
24. Ahmed, M., M. Abdel-Aty, and R. Yu. A Bayesian Updating Approach for Real-Time Safety Evaluation Using AVI Data. *Transportation Research Record,Journal of the Transportation Research Board*, Vol. 2450, 2012.
25. Wang, L., M. Abdel-Aty, Q. Shi, and J. Park. Real-Time Crash Prediction for Expressway Weaving Segments. *Transportation Research Part C: Emerging Technologies*, Vol. 61, 2015, pp. 1–10. https://doi.org/10.1016/j.trc.2015.10.008.
26. Xu, C., P. Liu, W. Wang, and Z. Li. Evaluation of the Impacts of Traffic States on Crash Risks on Freeways. *Accident; analysis and prevention*, Vol. 47, 2012, pp. 162–71. https://doi.org/10.1016/j.aap.2012.01.020.
27. Xu, C., W. Wang, and P. Liu. A Genetic Programming Model for Real-Time Crash Prediction on Freeways. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 14, No. 2, 2013, pp. 574–586. https://doi.org/10.1109/TITS.2012.2226240.
28. Lin, L., Q. Wang, and A. W. Sadek. A Novel Variable Selection Method Based on Frequent Pattern Tree for Real-Time Traffic Accident Risk Prediction. *Transportation Research Part C: Emerging Technologies*, Vol. 55, 2015, pp. 444–459. https://doi.org/10.1016/j.trc.2015.03.015.
29. Peng, Y., M. Abdel-Aty, Q. Shi, and R. Yu. Assessing the Impact of Reduced Visibility on Traffic Crash Risk Using Microscopic Data and Surrogate Safety Measures. *Transportation Research Part C: Emerging Technologies*, Vol. 74, No. January 2008, 2017, pp. 295–305. https://doi.org/10.1016/j.trc.2016.11.022.
30. Basso, F., L. J. Basso, F. Bravo, and R. Pezoa. Real-Time Crash Prediction in an Urban Expressway Using Disaggregated Data. *Transportation Research Part C: Emerging Technologies*, Vol. 86, No. January 2017, 2018, pp. 202–219. https://doi.org/10.1016/j.trc.2017.11.014.
31. Katrakazas, C., M. Quddus, and W. H. Chen. A Simulation Study of Predicting Real-Time Conflict-Prone Traffic Conditions. *IEEE Transactions on Intelligent Transportation Systems*, 2017, pp. 1–12. https://doi.org/10.1109/TITS.2017.2769158.
32. Katrakazas, C. Developing an Advanced Collision Risk Model for Autonomous Vehicles. PhD Thesis, Loughborough University, 2017.
33. Xu, C., 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, 2015, pp. 28–35. https://doi.org/10.1080/15389588.2014.909036.
34. Abdel-Aty, M., N. Uddin, A. Pande, F. Abdalla, and L. Hsia. Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression. *Transportation Research Record*, Vol. 1897, No. 1, 2004, pp. 88–95. https://doi.org/10.3141/1897-12.
35. Xu, C., W. Wang, P. Liu, R. Guo, and Z. Li. Using the Bayesian Updating Approach to Improve the Spatial and Temporal Transferability of Real-Time Crash Risk Prediction Models. *Transportation Research Part C: Emerging Technologies*, Vol. 38, 2014, pp. 167–176. https://doi.org/10.1016/j.trc.2013.11.020.
36. Abdel-Aty, M., and A. Pande. Identifying Crash Propensity Using Specific Traffic Speed Conditions. *Journal of safety research*, Vol. 36, No. 1, 2005, pp. 97–108. https://doi.org/10.1016/j.jsr.2004.11.002.
37. Pande, A., and M. Abdel-Aty. Assessment of Freeway Traffic Parameters Leading to Lane-Change Related Collisions. *Accident Analysis and Prevention*, Vol. 38, No. 5, 2006, pp. 936–948. https://doi.org/10.1016/j.aap.2006.03.004.
38. Yu, R., and M. Abdel-Aty. Utilizing Support Vector Machine in Real-Time Crash Risk Evaluation. *Accident; analysis and prevention*, Vol. 51, 2013, pp. 252–9. https://doi.org/10.1016/j.aap.2012.11.027.
39. Xu, C., W. Wang, and P. Liu. Identifying Crash-Prone Traffic Conditions under Different Weather on Freeways. *Journal of Safety Research*, Vol. 46, 2013, pp. 135–144. https://doi.org/10.1016/j.jsr.2013.04.007.

40. Sun, J., and J. Sun. A Dynamic Bayesian Network Model for Real-Time Crash Prediction Using Traffic Speed Conditions Data. *Transportation Research Part C: Emerging Technologies*, Vol. 54, 2015, pp. 176–186. https://doi.org/10.1016/j.trc.2015.03.006.
41. Yu, R., and M. Abdel-Aty. Analyzing Crash Injury Severity for a Mountainous Freeway Incorporating Real-Time Traffic and Weather Data. *Safety Science*, Vol. 63, 2014, pp. 50–56. https://doi.org/10.1016/j.ssci.2013.10.012.
42. Xu, C., P. Liu, W. Wang, and X. Jiang. Development of a Crash Risk Index to Identify Real Time Crash Risks on Freeways. *KSCE Journal of Civil Engineering*, Vol. 17, No. 7, 2013, pp. 1788–1797. https://doi.org/10.1007/s12205-013-0353-6.
43. Pande, A. Estimation of Hybrid Models for Real-Time Crash Risk Assessment on Freeways.PhD Thesis, University of Central Florida, 2005.
44. Pande, A., and M. Abdel-Aty. Assessment of Freeway Traffic Parameters Leading to Lane-Change Related Collisions. *Accident; analysis and prevention*, Vol. 38, No. 5, 2006, pp. 936–48. https://doi.org/10.1016/j.aap.2006.03.004.
45. Vogt, A., and J. G. Bared. *Accident Models for Two-Lane Rural Roads: Segment and Intersections - Report FHWA-RD-98-133*. 2008.
46. Sargent, D. J. Comparison of Artificial Neural Networks with Other Statistical Approaches. *Cancer*, Vol. 91, No. S8, 2001, pp. 1636–1642. https://doi.org/10.1002/1097-0142(20010415)91:8+<1636::AID-CNCR1176>3.0.CO;2-D.
47. Holland, J. H. Genetic Algorithms - Computer Programs That “evolve” in Ways That Resemble Natural Selection Can Solve Complex Problems Even Their Creators Do Not Fully Understand. *Scientific American*. 66–72.
48. Lv, Y., S. Tang, and H. Zhao. Real-Time Highway Traffic Accident Prediction Based on the K-Nearest Neighbor Method. *2009 International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2009*, Vol. 3, 2009, pp. 547–550. https://doi.org/10.1109/ICMTMA.2009.657.
49. Theofilatos, A. *An Advanced Multi-Faceted Statistical Analysis of Accident Probability and Severity Exploiting High Resolution Traffic and Weather Data*. PhD Thesis,National Technical University of Athens, 2015.
50. Theofilatos, A., G. Yannis, C. Antoniou, A. Chaziris, and D. Sermpis. Time Series and Support Vector Machines to Predict Powered-Two-Wheeler Accident Risk and Accident Type Propensity: A Combined Approach. *Journal of Transportation Safety and Security*, Vol. 9962, No. March, 2017, pp. 1–20. https://doi.org/10.1080/19439962.2017.1301611.
51. Fountas, G., M. T. Sarwar, P. C. Anastasopoulos, A. Blatt, and K. Majka. Analysis of Stationary and Dynamic Factors Affecting Highway Accident Occurrence: A Dynamic Correlated Grouped Random Parameters Binary Logit Approach. *Accident Analysis and Prevention*, Vol. 113, No. February 2017, 2018, pp. 330–340. https://doi.org/10.1016/j.aap.2017.05.018.
52. Breiman, L. Bagging Predictors. *Machine Learning*, Vol. 24, No. 421, 1996, pp. 123–140. https://doi.org/10.1007/BF00058655.
53. Ho, T. K. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, 1998, pp. 832–844. https://doi.org/10.1109/34.709601.
54. Breiman, L. Random Forests. *Machine learning*, Vol. 45.1, 2001, pp. 5–32. https://doi.org/10.1017/CBO9781107415324.004.
55. Verikas, A., A. Gelzinis, and M. Bacauskiene. Mining Data with Random Forests: A Survey and Results of New Tests. *Pattern Recognition*, Vol. 44, No. 2, 2011, pp. 330–349. https://doi.org/10.1016/j.patcog.2010.08.011.
56. Christodoulou, C., and M. Georgiopoulos. *Applications of Neural Networks in Electromagnetics*. Artech House, Inc., Norwood, MA, USA, 2001.
57. Tomek, I. An Experiment with the Edited Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 6, No. 6, 1976, pp. 448–452. https://doi.org/10.1109/TSMC.1976.4309523.

58. Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority over-Sampling Technique. *Journal of Artificial Intelligence Research*, Vol. 16, 2002, pp. 321–357. https://doi.org/10.1613/jair.953.
59. Batista, G. E. A. P. A., R. C. Prati, and M. C. Monard. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explor. Newsl.*, Vol. 6, No. 1, 2004, pp. 20–29. https://doi.org/10.1145/1007730.1007735.
60. Wilson, D. L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 2, No. 3, 1972, pp. 408–421. https://doi.org/10.1109/TSMC.1972.4309137.
61. Guan, D., W. Yuan, Y. K. Lee, and S. Lee. Nearest Neighbor Editing Aided by Unlabeled Data. *Information Sciences*, Vol. 179, No. 13, 2009, pp. 2273–2282. https://doi.org/10.1016/j.ins.2009.02.011.
62. Yannis, G., J. Golias, C. Antoniou, S. Vardaki, P. Papantoniou, D. Pavlou, S. Espié, G. Kalisperakis, S. G. Papageorgiou, G. Tsivgoulis, A. Bonakis, N. Andronas, I. Papatriantafyllou, A. Liozidou, D. Kontaxopoulou, A. Economou, and M. Kosmidis. *Distract: Causes and Impacts of Driver Distraction: A Driving Simulation Study, Deliverable 4: Driving Simulator Experiment*. 2014.
63. Pavlou, D., E. Papadimitriou, P. Papantoniou, G. Yannis, and S. G. Papageorgiou. The Impact of Cognitive Impairments on Accident Risk. 7th International Conference on ESAR „Expert Symposium on Accident Research“. Reports on the ESAR-Conference 2016 at Hannover Medical School.
64. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, Vol. 12, 2012, pp. 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2.
65. Lemaitre, G., F. Nogueira, and C. K. Aridas. Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *CoRR*, Vol. abs/1609.0, 2016, pp. 1–5.
66. Karlaftis, M. G., and E. I. Vlahogianni. Statistical Methods versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights. *Transportation Research Part C: Emerging Technologies*, Vol. 19, No. 3, 2011, pp. 387–399. https://doi.org/10.1016/j.trc.2010.10.004.