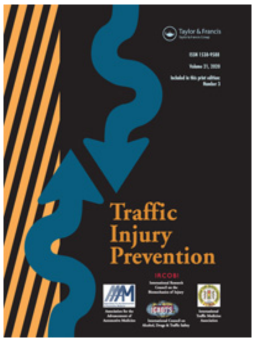

Taylor & Francis
Taylor & Francis Group

Traffic Injury Prevention

ISSN: 1538-9588 (Print) 1538-957X (Online) Journal homepage: https://www.tandfonline.com/loi/gcpi20

Class-imbalanced crash prediction based on real-time traffic and weather data: A driving simulator study

Zouhair Elamrani Abou Elassad, Hajar Mousannif & Hassan Al Moatassime

To cite this article: Zouhair Elamrani Abou Elassad, Hajar Mousannif & Hassan Al Moatassime (2020) Class-imbalanced crash prediction based on real-time traffic and weather data: A driving simulator study, Traffic Injury Prevention, 21:3, 201-208, DOI: 10.1080/15389588.2020.1723794

To link to this article: https://doi.org/10.1080/15389588.2020.1723794

TRAFFIC INJURY PREVENTION
2020, VOL. 21, NO. 3, 201–208
https://doi.org/10.1080/15389588.2020.1723794

Class-imbalanced crash prediction based on real-time traffic and weather data: A driving simulator study

Zouhair Elamrani Abou Elassad[a], Hajar Mousannif[a], and Hassan Al Moatassime[b]

[a]LISI Laboratory, Computer Science Department, FSSM, Cadi Ayyad University, Marrakesh, Morocco; [b]OSER Research Team, Computer Science Department, FSTG, Cadi Ayyad University, Marrakesh, Morocco

ABSTRACT

Objective: Crash occurrence prediction has been of major importance in proactively improving traffic safety and reducing potential inconveniences to road users. Conventional statistical crash prediction models frequently suffer from severe data quality issues and require a significant amount of historical data. On the other hand, even though machine learning (ML) based algorithms have proven to be powerful in predicting future outcomes in different fields of applications, they likely fail to provide satisfactory results unless a tuning parameter approach is conducted. The main objective of this article is to develop real-time crash prediction models that will potentially be employed within traffic management systems.

Methods: In this study, two highly optimized data-driven models for crash occurrence prediction have been designed based on the popular machine learning techniques, Support Vector Machine (SVM) and deep neural network Multilayer Perceptron (MLP). To ensure that the proposed algorithms produce robust and stable performance, the optimal scheme for models' construction has been thoroughly examined and discussed. Additionally, the further boost of models' performance requires the systemic assessment of crash strongest precursors within the driver-vehicle-environment triptych. Therefore, three categories of features, including driver input responses, vehicle kinematics and weather conditions, were measured during the execution of various driving tasks performed on a desktop driving simulator. Moreover, since crash events typically occur in rare instances tending to be underrepresented in the dataset, an imbalance-aware strategy to overcome the issue was adopted using the Synthetic Minority Oversampling TEchnique (SMOTE).

Results: The results show that MLP exhibited the best performing prediction results, most particularly, in clear, overcast and snow conditions, in which MLP recall values were above 94%. Higher F1-score values were achieved in overcast and rain weather by MLP and snow conditions by SVM; whereas over 90% of G-mean levels were obtained under fog and rain conditions for MLP and snow condition for SVM.

Conclusion: The findings provide new insights into crash events forecasting and may be used to promote enforcement efforts related to designing crash avoidance/warning systems that enhance the effectiveness of the system's application based on driver input and vehicle kinematics under various weather conditions.

ARTICLE HISTORY
Received 28 October 2019
Accepted 27 January 2020

KEYWORDS
Crash prediction; driving simulator; support vector machine; multilayer perceptron; machine learning; SMOTE

Introduction

Road traffic accidents are known to be one of the most major concerns and threatening problems that encounters societies nowadays, resulting in many health issues, economic losses and fatalities. The World Health Organization (WHO 2017) reports that 1.35 million people die in road traffic crashes every year, and a further 20-50 million are injured or disabled worldwide. Hence, comprehending under what circumstances crashes occur and which factors rise the likelihood of a car accident, would have a sizeable impact on developing effective policy interventions in order to avoid accidents from happening. Reliable crash occurrence forecasting and proactive safety analysis are undeniably of great interest and necessity.

A vast majority of studies have been conducted on traffic accident evaluation and prevention. Crash occurrence is a complex mechanism, influenced by multiple contributing factors such as the driver state and environmental factors (Aljanahi, Rhodes, and Metcalfe 1999). Hazardous traffic conditions and unsafe driving behaviors have been examined in numerous previous studies in order to characterize road crashes and develop efficient real-time traffic management strategies (Ahmed and Abdel-Aty 2012; Shi et al. 2018). Although the previous studies provide useful insights into promoting positive safety practices, it is crucial to note that in the process of crash prediction various modeling techniques result in different performance measures. Within this context, machine learning models have proven to supersede

CONTACT Zouhair Elamrani Abou Elassad z.elamrani@uca.ma LISI Laboratory, Computer Science Department, FSSM, Cadi Ayyad University, Marrakesh 40000, Morocco

Associate Editor Alessandro Calvi oversaw the review of this article.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/15389588.2020.1723794

statistical analysis in predicting forthcoming events and have reported satisfying results in many transportation systems (Elamrani Abou Elassad et al. 2020; Lee, Derrible, and Pereira 2018; Mousannif et al. 2016). The Support Vector Machines (SVM) and Deep Neural Networks (DNN) are ones of the most substantial machine learning techniques that have been used for crash occurrences prediction (Basso et al. 2018; Li et al. 2016; Li et al. 2018; Theofilatos, Chen, and Antoniou 2019; Yu and Abdel-Aty 2013). The potential of adopting SVM in assessing safety performance measures for vehicle crashes depicted a better goodness-of-fit comparing to negative binomial models (Li et al. 2008). It was argued that SVM model can handle small data sizes, with a great capability in producing fewer over-fitting issue and better generalization abilities (Yu and Abdel-Aty 2013). On the other hand, Deep Neural Networks have gain their popularity owing to their excellence in various complex tasks; they have been gradually recognized for their ability to learn data representations in both supervised and unsupervised settings along with parallel processing, fault tolerance, and their efficiency to generalize to unseen data samples using hierarchical representations (Basu et al. 2018). Reflecting on literature related to the field crash prediction, techniques like SVM and DNN are deemed to be prominent and effective in crash prediction systems based on their powerful theoretical grounding. However, all these machine learning algorithms face a great challenge in order to provide optimal performance results, therefore, optimization strategy of parameters have been conducted. Based on the previous research, it is proven that in order to get improved performance metrics in SVM model, penalty factor (C), and kernel parameter (γ) are considered to be optimized. SVM optimization was carried out using Grid Search method which is one of the most widely used methods in and it has been proven as an efficient way to the model's hyperparameters (Hsu, Chang, and Lin 2003). Likewise, DNN optimization can be depicted by tuning the number of layers, input and hidden neurons, weights, etc. Trial and error strategy along with the dropout regularization method (Srivastava et al. 2014) and the early stop approach were used to this end. K-fold cross-validation technique was adopted to evaluate the classification performance. It has been recognized for its susceptibility to yield minimal bias and variance in contrast with the other validation methods, including the leave-one-out method (Kohavi 1995).

The required data for crash prediction systems can be collected across different experiment forms, namely naturalistic driving studies as well as field driving studies and driving simulator studies (Elamrani Abou Elassad and Hajar 2019). Choosing Simulator driving studies (SDS) in the field of transportation research has been growing in recent years as they try to simulate driving conduct in a safe environment, with the major benefit of possessing full empirical control over conditions and the ability to investigate multiple design structures. As such, vehicle telemetry such as speed and drift angle, driver input like steering wheel position and throttle pedal position, in addition to multiple weather scenarios namely overcast, fog, rain and snow seasons have been simulated and recorded during the experiments. Driver behavioral responses along with vehicle kinematics have been shown to have a vital impact on safe driving and on the identification of crash and near-crash events (Ba et al. 2016; McDonald et al. 2018; Michaels et al. 2017; Perez et al. 2017). On the other hand, it was found that more than 1.25 million accidents are caused yearly due to weather conditions (21% of all vehicle crashes), leading to about 418,000 injuries (19% of crash injuries), and nearly 5000 casualties (16% of all casualties) (FHWA 2016). Weather variables have been investigated in crash prediction strategies and found to be highly influential (Abdel-Aty and Pemmanaboina 2006; Madanat and Liu 1995; Wang, Shi, and Abdel-Aty 2015). Even though there are several studies that examined the impact of weather conditions in analyzing crash events, research investigating the prediction of crash occurrences in multiple weather conditions is relatively limited. Another instrumental factor in the prediction of crash events is the proportion of crash and non-crash instances in the dataset. Crash related observations usually produce imbalanced data sets since the target classes are not equally represented. Handling the issue of imbalanced dataset is a challenging procedure for which scholars are seeking to enhance and harness different technologies. To this end, Synthetic Minority Oversampling Technique (SMOTE), deemed as one of the most powerful re-sampling algorithms, was presented by (Chawla et al. 2002) to solve the imbalance issue by producing synthetic instances from the minor class. An extensive research have proven that SMOTE has a better efficiency than under-sampling and over-sampling techniques (Batuwita and Palade 2013; Kaur and Gosain 2018; Nguyen, Cooper, and Kamei 2011).

Based on the above stated literature, crash occurrences prediction studies still meet a few challenges or suffer from limited efficiency. There is still a need to construct more tuned algorithms to competently predict crash events. Variable selection procedure was adopted based on Random Forest algorithm in order to examine the strongest precursors of crash events and to capture the impacts of explanatory features on the overall prediction accuracy. Optimized MLP (Multilayer perceptron) and SVM machine learning techniques were developed to achieve better performance results. To the best of our knowledge, little to no research has directly constructed a thorough treatment of the proposed models in predicting crash occurrences accounting for the SMOTE imbalance-aware learning strategy using driver input characteristics, vehicle telemetry and weather conditions. Furthermore, minimal work has been directed to quantify crash events in real-time weather data. As such, the objectives of this paper are twofold: (1) to expand the current knowledge of crash occurrence investigation by adopting of real-time information incorporating various weather conditions (i.e., clear, overcast, fog, rain, snow), vehicle telemetry (i.e., speed, yaw angle, etc.) and driver input features (i.e., accelerator/brake pedal position, steering angle) on crash outcome prediction. (2) to adopt an imbalance-learning strategy based on the SMOTE technique in order to develop optimized machine learning models for crash prediction.

The remainder of this study is organized as follows. First, descriptions of the driving simulator and experimental protocol with data analysis are provided. Second, in the methodology section, the construction of modeling techniques is presented in details. Next, the results are reported and interpreted. Finally, conclusions with future scopes of the present study are offered.

Driving simulator experiment

Participants and apparatus

A total of 62 volunteers (43 males and 19 females) between the ages of 20 and 51 (M = 40.25; SD = 2.20) participated in the study. All participants had a full driver's license and had been driving for at least a year. Average years of driving experience ranged from 1 to 17 years (M = 10.45; SD = 6.78) with an average hour of driving per day ranging from 1 to 6 h (M = 3.20, SD = 2.39). All were in a good health, and had (corrected to) normal vision. In reference to the provided information about the experiment's general intentions, all participants were naïve to the purpose of the study and gave informed consent form about data recording of their driving performance. The study was carried out using a fixed-based driving simulator located at the University of Cadi Ayyad (UCA) facility. Simulator driving studies hold a major advantage of simulating conduct in a safe environment with a full experimental control over driving conditions including all types of weather, terrain, and traffic (Elamrani Abou Elassad and Hajar 2019). Surely, it would be very dangerous to carry out trials on real road environment. The driving simulation was run through the Project Cars 2 simulator by (Slightly Mad Studios) using a Logitech® G27 Racing Wheel set (steering wheel, accelerator pedal, and brake pedal) with the adjustable Logitech Evolution® Playseat, simulations were conducted with automatic gear selection, thus gear shifter was not needed. Figure 1 illustrates the hardware setup.

Simulation scenario and procedure

The driving scenario was performed in daylight and under five different sequential weather conditions (clear, overcast, fog, rain and snow) and aimed to simulate various intricacies and aspects that real-world driving entails in order to explore the impact of the factors on driving behavior and to collect enough raw data before the crash. The adopted route layout was the same for all the participants; each participant completed 5 drive sessions in the simulator, the first visit in each drive was a practice session; then drivers underwent the simulation once at each of the five weather conditions (clear, overcast, fog, rain and snow). The idea behind the scenario was to instruct participants to drive as they normally would as the driving trials consisted of a set of maneuvers mimicking a standard on-road evaluation while traffic flow density was kept constant during each experiment. Each participant started the simulation from the same start point and navigated the vehicle to arrive to the end point. All subjects were requested to obey the rules of the road like the lights and traffic signs, representative buildings and landmarks along with traffic lights and stop signs were included in the driving environment. In the trials' scenarios, when ambient road users intruded driver's pathway and triggered a conflict leading to a potential collision, the traffic conflict created in each scenario was an observable situation in which two or more road users approached each other in space and time to such an extent that there was a risk of collision. It is noteworthy that in studies of crash prediction, the crash events are generally unexpected and occur rarely, as such, this adopted type of road test contains most of the elements typically used to analyze crash events (e.g., ability to make right turns, change lanes, use signals, etc.). This set of weather and terrain characteristics serve to provide different levels of difficulty while maneuvering the vehicle along the driving route. The drivers were appointed to a quiet laboratory to virtually drive the vehicle; each participant navigated experimental session drives on a virtual two-lane urban road of 19.25 km length which required about 13 min to complete when the speed limits are kept. More details about the adopted scenario and procedure can be found in the Appendix (online supplement).

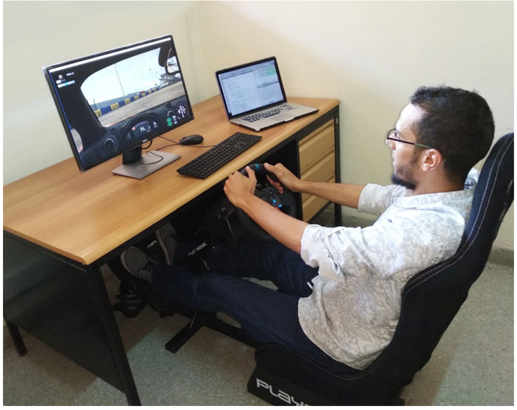

Figure 1. Experimental setup of the desktop driving simulator at UCA.

Collected data

Data were continuously recorded throughout each drive with a sampling frequency of 60 Hz through the User Datagram Protocol (UDP). The simulator collects records of driver inputs (e.g., throttle/brake pedal position, steering wheel position), vehicle dynamics (e.g., speed, yaw angle) and environmental data such as the weather season in which the season attribute (see Figure A3 in the Appendix, online supplement) prevails the weather condition at the time of the accident namely overcast, fog, rain or snow. For instance, when the drivers' vehicle entered a risky scenario, drivers released the gas pedal and slowed down the speed. After the hazard vehicle emerged in the visual field, the driver first detected it and then started the brake reaction subsequently in an attempt to decelerate the vehicle. On another hand, features such as the yaw angle can depict the stability of the vehicle in turning maneuvers and the ability to avoid the motion of skidding especially in rain and snow weather conditions. These are three categories of metrics that may affect traffic safety. The dependent variable is crash occurrence, coded as a binary variable with a value of 1 if a crash was identified and 0 if not. With Reference to the previous research of the intervention time (Werneke and Vollrath 2013; Yan, Zhang, and Ma 2015), we retrieved the 12 s length data segments, from 16 s to 4 s prior to the crashes, as the crash data to validate the timeliness of the suggested prediction strategy. In parallel, information segments of 12 s duration were randomly extracted from all 62 drivers' raw data as the non-crash data, which did not overlap with any crash instances. A summary of the grouping and definitions for all the features acquired during the driving simulations is presented in the Appendix (Table A1, online supplement).

Methodology

The main objective of this study is to develop a model to predict crash occurrences by considering the most relevant features and implementing the most effective machine leaning techniques. Crash prediction is a binary classification model aiming to determine whether a driver will have an accident or not. As such, a three-stage process is considered: pre-processing, variable selection, and prediction model construction. Since accident event probability is deemed to be generally small which will lead to have a highly imbalanced classification problem; pre-processing of the data set is conducted to handle the problem of imbalanced class distribution. Variable selection is used to remove redundant or irrelevant variables in order to find a set of relevant features that better describe our data, and ideally, result in a more robust prediction performance. Then, two different prediction models are constructed, a support vector machine (SVM) and a Multilayer Perceptron (MLP) to predict crash occurrences and to evaluate their performance metrics. The comprehensive process for models' construction and optimization is outlined in the Appendix (the Building Prediction Models section, online supplement).

Pre-Processing

The ratio of the crash data and non-crash data produced an unbalanced data set as outcome classes were not equally represented, a prevalent characteristic in the crash occurrence evaluation. Such imbalances cause a bias toward the majority class, since modeling classifiers prioritize the class with the higher number of observations causing an over-prediction of the this class (Fernández, del Jesus, and Herrera 2009). Pre-processing aims to resolve this issue by balancing class distribution in the data set. In this paper, the synthetic minority over-sampling technique (SMOTE) was adopted, presented by (Chawla et al. 2002). The SMOTE technique creates synthetic minority instances based at random intervals between existing minority cases rather than duplicating existing minority cases. For more information, see the Pre-Processing section in the Appendix.

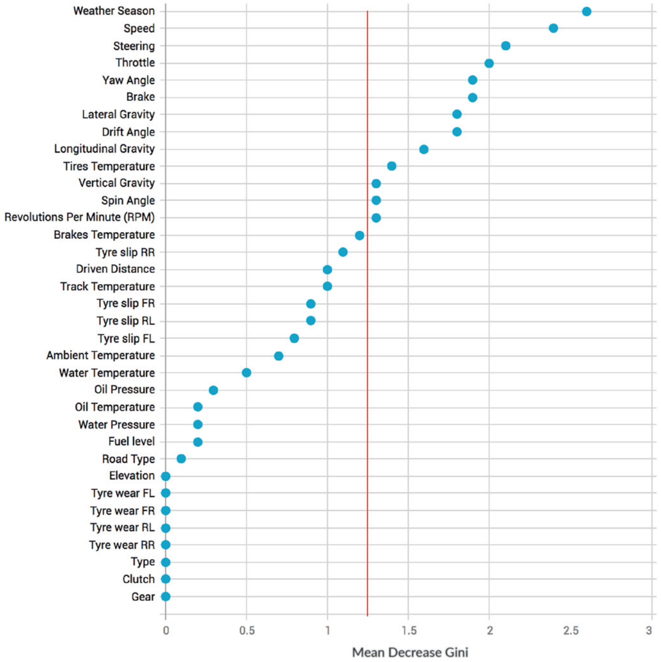

Figure 2. Variable importance ranking using Gini impurity index.

Variable selection

After pre-processing, reducing the dimensionality of the input variables through variable selection is an important step for classification models as some of these features may correlate with each other or not have any considerable effects on crash occurrence. That is, including a large number of features without any process of variable selection may lead to increased error of estimates (Kan et al. 2019). Thus, it becomes substantial to inspect our data to determine what are the variables that appear to be strong precursors of car accidents. Random Forest (RF) models have been widely adopted for variable selection in crash risk analyses (Siddiqui, Abdel-Aty, and Huang 2012; Yu et al. 2019). Based on the RF Gini index measure, which is used to evaluate variable importance, a total of 35 variables were culled to 13 by the selection process as can be seen in Figure 2. Further information are provided in the Variable Selection section in the Appendix (online supplement). The grouping for all the selected candidate factors is as follows:

- Vehicle Kinematics:
  - Speed
  - Lateral Gravity
  - Longitudinal Gravity
  - Vertical Gravity
  - Yaw Angle
  - Drift Angle
  - Spin Angle
  - Revolutions Per Minute (RPM)
- Driver Input:
  - Throttle
  - Brake
  - Steering
- Environmental Conditions
  - Weather Season

Prediction models

Support vector machine

Developed by (Vapnik 1995), support vector machine (SVM) is a non-probabilistic binary linear classifier that can be used to solve a classification problem by constructing optimal separating hyperplane in a manner that the margin is maximized so that SVM has good generalization ability. A basic structure of a simple SVM is illustrated in the Appendix (Figure A5, online supplement). A set of related supervised learning methods used for prediction and regression based on the statistical learning theory and structural risk minimization which are the theoretical foundations for the SVMs learning algorithms. It has been proven that SVM is highly efficient and robust algorithm for binary classification problems, and it has been found to demonstrate proportionate or superior performance than other statistical and machine learning methods (Hsu, Chang, and Lin 2003; Kecman 2005). For more details about the SVM method, see its dedicated section in the Appendix.

Multi-layer perceptron

Artificial neural networks (ANN) are abstract computational techniques inspired by biological neural networks and are efficient and applicable for predicting the relation between dependent and independent parameters. The performance of ANN prediction is highly affected by its structure which is comprised of an input layer, hidden layers and an output layer; each layer contains a group interconnected processing units called neurons or nodes that are effective in processing enormous parallel computations and knowledge representation (Basheer and Hajmeer 2000). The deep learning multi-layer perceptron (MLP) is one of the most performant types of ANN that was selected to improve the classification prediction performance. For more details about the MLP method, see its dedicated section in the Appendix (online supplement).

Model quality assessment

A variety of frequently used performance measures are used to evaluate the quality of the classification models. All of them are gained from the confusion matrix, a two-dimensional contingency table that shows the agreement between predicted and actual class and deemed to be the basis of the predictability power of the model (Davoudi Kakhki, Freeman, and Mosher 2019). The confusion matrix (Table A2 in the Appendix, online supplement) presents performance evaluation measure by using a confusion matrix where the True Positive (TP) indicates the number of crash occurrences correctly classified, and False Positive (FP) indicates the number of non-crash events incorrectly classified as crash-events. False Negative (FN) indicates the number of crash events incorrectly classified as non-crash events, and True Negative (TN) indicates the number of non-crash occurrences correctly classified. In this work, the accuracy metric along with Recall, Precision, G-mean, F1-score and AUC measures have been employed, more information about the models' evaluation strategy can be consulted in the Model Quality Assessment section in the Appendix.

Experimental results

In an effort to demonstrate the validity of the classifiers' evaluation, parameter optimization for each of two classifiers SVM and ANN was carried out. The complete details of models' construction can be found in the Building Prediction Models section in the Appendix. In order to cope with imbalanced dataset, SMOTE was employed to rebalance the training sets; when dealing with imbalanced datasets, accuracy may suffer due to bias toward the majority class (Haixiang et al. 2017). Therefore, it is critically substantial to select the appropriate measurement metrics to assess classifier efficiency and guide model learning. There are multiple performance evaluation metrics that take class distributions into consideration, such as the AUC, Recall, G-mean and F1-score. The average prediction of these performance measures based on the cross-validation results for the different classification models are shown in Table 1.

Based on the results, it appears that in clear weather, except for precision and accuracy in which SVM exhibited high performance with values of 85.03% and 92.24% respectively, while MLP higher values in the rest of the assessment metrics with 95.12% recall, 89.35% F1-score, 93.95% G-mean and 94.24% AUC. In overcast weather, a higher 92.13% G-mean was obtained with SVM; MLP achieved superior values in the rest with 89.06% precision, 95.5% recall, 92.17% F1-score, 93.17% AUC and 95.29% accuracy. When it comes to fog conditions, except for recall and G-mean in which MLP achieved higher values with 94.01% and 95.34% respectively, SVM outperformed MLP in what remains with 92.80% precision, 89.46% F1-score, 94.82% AUC and 94.59% accuracy. As for rainy weather conditions, in terms of recall and AUC, SVM exhibited the best results with values of 91.58% and 92.20% respectively, whereas MLP attained better precision with a value of 93.77%, as well as high F1-score, G-mean and accuracy with values of 91.54%, 95.04% and 94.02% appropriately. Lastly for the snowy weather conditions, increased values of precision, F1-score and G-mean were obtained using SVM with values of 89.51%, 91.64% and 94.82% respectively, while MLP achieved 95.07% recall, 94.90% AUC and 94.16 accuracy. The overall comparison of modeling performance during the different weather conditions is presented in the Appendix (Figure A7, online supplement).

As can be seen, when comparing the recall and precision measures, it is clear that the recall levels achieved by the two models are, in general, much higher than precision. The

Table 1. Performance metrics for the crash events using MLP and SVM during different weather conditions.

Model	Weather	Precision	Recall	F1-score	G-mean	AUC	Accuracy
MLP	Clear	84.24	95.12	89.35	93.95	94.24	90.07
SVM		85.03	91.04	87.93	92.07	92.12	92.24
MLP	Overcast	89.06	95.5	92.17	90.44	93.17	95.29
SVM		82.11	95.1	88.13	92.13	87.62	89.70
MLP	Fog	84.20	94.01	88.83	95.34	93.00	93.15
SVM		92.80	86.36	89.46	93.22	94.82	94.59
MLP	Rain	93.77	89.42	91.54	95.04	90.71	94.02
SVM		86.19	91.58	88.80	92.09	92.20	90.35
MLP	Snow	84.96	95.07	89.73	92.10	94.90	94.16
SVM		89.51	93.88	91.64	94.82	91.62	93.11

higher the recall, the more accurate the prediction result as it represents the correct crash prediction overall actual crash records. Put differently, it depicts the proportion of crash occurrences that were correctly predicted by the models. Most of the highest recall values were obtained with MLP with performance over 90%, especially in clear, overcast and snow conditions, in which MLP recall values were beyond 94%. Conventionally, precision and recall metrics have an inherent tradeoff as one comes at the cost of the other. Thus, F1 score is a special measure that conveys the balance between the precision and recall in order to find an effective and efficient tradeoff, as evidenced, good measures were obtained of F1-score with a minimum of 87%, with superior results (above 90%) in both overcast and rain weather for MLP and snow conditions for SVM. In all weather conditions, the G-mean metric values, which has been found to yield more accurate performance measures when the underlying data is impacted with imbalance, achieved levels over 90%, most particularly in fog and rain conditions with G-mean over 94% for MLP and SVM for snow conditions. The average performance measures for MLP and SVM compared with three other adopted models: Naïve Bayes (NB), Hidden Markov Model (HMM) as well as Logistic Regression (LR) are depicted in Table 2. The MLP model appears to be the best performing classifier in terms of average performance for all the modeling metrics. This clearly indicates that, in this context, the use of MLP is preferable. To further explore the findings visually, the aforementioned results are presented in Figure 3.

A comparison between the acquired findings and other conventional crash prediction studies is presented in the Appendix (Table A3, online supplement). Several modeling techniques that have been compared based on weather conditions, feature selection, class imbalance and the best level of performance in terms of the adopted evaluation metrics.

Table 2. Performance metrics for the crash events using different modeling techniques.

Model	Precision	Recall	F1-score	G-mean	AUC	Accuracy
MLP	87.25	93.82	90.33	93.37	93.20	93.34
SVM	87.13	91.59	89.19	92.87	91.68	92.00
NB	80.09	76.78	78.27	82.15	80.44	80.83
HMM	72.00	83.19	79.84	85.30	78.97	84.64
LR	67.11	75.53	71.17	70.22	68.09	77.37

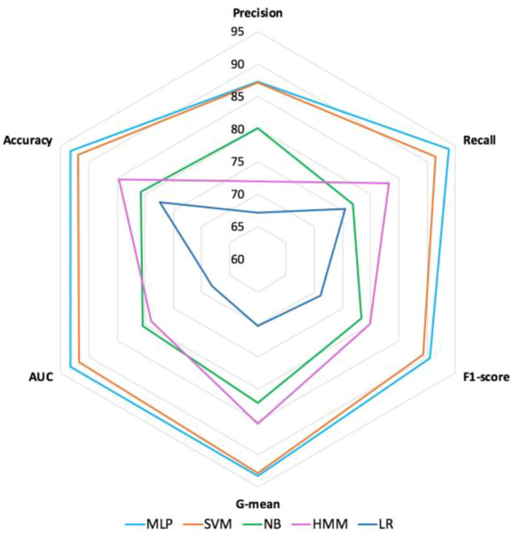

Figure 3. Average performance overview for classification models.

Discussion and conclusion

Traffic safety improvement is deemed to be a major concern worldwide. Since traffic crashes result in countless health issues, economic losses and fatalities, a more comprehensive analysis that aims to reduce traffic crashes and enhance traffic safety using effective and highly accurate real-time crash prediction models is required. The purpose of this study is to develop crash occurrence prediction algorithms using machine learning techniques and to identify crash strongest precursors which mean the driving conditions leading to crash events.

In this paper, Support Vector Machine (SVM) and the deep learning algorithm Multilayer Perceptron (MLP) models, considered as one of the most substantial machine learning techniques that have been used for traffic safety analysis, were proposed for crash events prediction. The methodology adopted in this work substantially differs from previous studies as it provides an overall assessment of crash strongest precursors using optimized models within the driver-vehicle-environment triptych; three categories of features, including driver input responses, vehicle kinematics and weather conditions, were measured during the execution of various driving tasks performed using a desktop driving simulator. Crash related observations usually produce imbalanced data sets as the target classes are not equally represented. Machine learning techniques are not very good at predicting less representative class in imbalanced datasets, especially if they are not properly optimized. Therefore, a parameter tuning strategy along with a data balancing task is needed as part of the data preprocessing phase. Within this contest, grid search for SVM and trial and error approach with the dropout regularization for MLP were conducted to examine the optimal model configurations. Moreover, to cope with imbalanced datasets, the SMOTE was used to rebalance the target training data.

The results of this study show that, if proper preprocessing procedure was applied, machine learning models are capable of predicting crash occurrence with high level of performance. The achieved recall values by the two models are, in general, much higher than precision, consequently, the constructed models are efficient in predicting the proportion of crash occurrences that were correctly predicted. Most of the highest recall values were obtained with MLP with performance over 90%, especially in clear, overcast and snow conditions, in which MLP recall values were beyond 94%. On another note, F1-score, which is a special measure that conveys the balance between the precision and recall in order to find an effective and efficient tradeoff, has exhibited during higher results in overcast and rain weather for MLP and snow conditions for SVM. When it comes to the G-mean metric values, which has been found to yield more accurate performance measures when the underlying data is impacted with imbalance, levels over 90% were achieved, most particularly in fog and rain conditions for MLP and SVM for snow conditions. From the usability standpoint, since the primary objective is to predict instances of a minority class, recall, F1 score and G-mean measures are more important than accuracy since a model could achieve a high accuracy while having a low recall. Finally, the MLP model appears to be the best performing classifier in terms of average performance for all the modeling metrics. This clearly indicates that, in this context, the use of MLP is preferable.

To be admitted, there exists some limitations that need to be addressed. Simulator studies provide a convenient and adjustable environment; however, the driving simulator is not an integral substitute for the real-world driving experiences. Indeed, the authenticity of the results acquired using a driving simulator rely on the tasks in question within the simulated environment. However, crash data during fairly similar conditions based on related research could be further processed. Further to this, although the above results have evidenced the performances of crash occurrence prediction with driver input, vehicle kinematics and weather conditions, additional complicated measures such as drivers' physiological and mental states measures could provide other insightful measures. Relevant noninvasive systems have been broadly adopted in driving behavior analysis such as wrist bands calculating Heart Rate Variability and Body Temperature which have been proven to be efficient workload assessment metric (Backs et al. 2003; Chen et al. 2017; Jha, Prakash, and Sagar 2018), as well as visual-based systems that gather data through cameras mounted in the vehicle (Henni et al. 2018; Ragab et al. 2014). Finally, potential future directions of this study may comprise broadening the predictive models to include ensemble methods and investigate various undersampling and oversampling techniques to handle class imbalance issue for the better prediction of crash events.

Acknowledgment

This research was jointly supported by (1) the Moroccan Ministry of Equipment, Transport and Logistics, and (2) the Moroccan National Center for Scientific and Technical Research (CNRST).

Funding

Centre National pour la Recherche Scientifique et Technique.

References

Ahmed MM, Abdel-Aty MA. 2012. The viability of using automatic vehicle identification data for real-time crash prediction. IEEE Trans Intell Transport Syst. 13(2):459–468. doi:10.1109/TITS.2011.2171052

Abdel-Aty MA, Pemmanaboina R. 2006. Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data. IEEE Transactions on Intelligent Transportation Systems 7(2): 167–74. doi:10.1109/TITS.2006.874710

Aljanahi AAM, Rhodes AH, Metcalfe AV. 1999. Speed, speed limits and road traffic accidents under free flow conditions. Accid Anal Prev. 31(1–2):161–68. doi:10.1016/S0001-4575(98)00058-X

Ba Y, Zhang W, Chan AHS, Zhang T, Cheng ASK. 2016. How drivers fail to avoid crashes: A Risk-Homeostasis/Perception-Response (RH/PR) framework evidenced by visual perception, electrodermal activity and behavioral responses. Transport Res Part F: Traffic Psychol Behav. 43:24–35. doi:10.1016/j.trf.2016.09.025

Backs RW, Lenneman JK, Wetzel JM, Green P. 2003. Cardiac measures of driver workload during simulated driving with and without visual occlusion. Human Factors. 45(4):525–38. doi:10.1518/hfes.45.4.525.27089

Basso F, Basso LJ, Bravo F, Pezoa R. 2018. Real-time crash prediction in an urban expressway using disaggregated data. Transp Res Part C: Emerg Technol. 86:202–19. doi:10.1016/j.trc.2017.11.014

Basu S, Mukhopadhyay S, Karki M, DiBiano R, Ganguly S, Nemani R, Gayaka S. 2018. Deep neural networks for texture classification—A theoretical analysis. Neural Networks. 97:173–82. doi:10.1016/j.neunet.2017.10.001

Basheer IA, Hajmeer M. 2000. Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods 43(1): 3–31. doi:10.1016/S0167-7012(00)00201-3

Basu S, Mukhopadhyay S, Karki M, DiBiano R, Ganguly S, Nemani R, Gayaka S. 2018. Deep neural networks for texture classification—A theoretical analysis. Neural Networks. 97:173–82. doi:10.1016/j.neunet.2017.10.001

Batuwita R, Palade V. 2013. Class imbalance learning methods for support vector machines. In: Imbalanced Learning. Hoboken, NJ: John Wiley & Sons, Inc. p. 83–99.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research. 16:321–57. doi:10.1613/jair.953

Chen L-l, Zhao Y, Ye P-f, Zhang J, Zou J-z. 2017. Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. Expert Syst Appl. 85:279–291. doi:10.1016/j.eswa.2017.01.040

Davoudi Kakhki F, Freeman SA, Mosher GA. 2019. Evaluating machine learning performance in predicting injury severity in agribusiness industries. Safety Science. 117: 257–62. doi:10.1016/j.ssci.2019.04.026

Elamrani Abou Elassad Z, Mousannif H. 2019. Understanding driving behavior: Measurement, modeling and analysis. In: Advances in Intelligent Systems and Computing. Vol. 5; p. 452–464. doi:10.1007/978-3-030-11928-7_41

Elamrani Abou Elassad Z, Mousannif H, Al H, Karkouch A. 2020. Engineering applications of artificial intelligence the application of machine learning techniques for driving behavior analysis: A conceptual framework and a systematic literature review. Eng Appl Artificial Intell. 87(March 2019):103312. doi:10.1016/j.engappai.2019.103312

Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. 2017. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications. 73:220–39. doi:10.1016/j.eswa.2016.12.035

Henni K, Mezghani N, Gouin-Vallerand C, Ruer P, Ouakrim Y, Vallières E. 2018. Feature selection for driving fatigue characterization and detection using visual- and signal-based sensors. Applied Informatics. 5(1):1–15. doi:10.1186/s40535-018-0054-9

Hsu CW, Chang CC, Lin CJ. 2003. A practical guide to support vector classification. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Jha V, Prakash N, Sagar S. 2018. Wearable anger-monitoring system. ICT Express. 4(4):194–98. doi:10.1016/j.icte.2017.07.002

Kan HJ, Kharrazi H, Chang H-Y, Bodycombe D, Lemke K, Weiner JP. 2019. Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults. PLOS ONE 14(3): e0213258. doi:10.1371/journal.pone.0213258

Kaur P, Gosain A. 2018. Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise. Singapore: Springer, p. 23–30.

Kohavi R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. 1137–1143 [accessed 2019 Jul 9]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.529.

Lee D, Derrible S, Pereira F C. 2018. Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling. Transport Res Record. 2672(49):101–112. doi:10.1177/0361198118796971

Li L, He S, Zhang J, Ran B. 2016. Short-term highway traffic flow prediction based on a hybrid strategy considering temporal-spatial information. J Adv Transp. 50(8):2029–40. doi:10.1002/atr.1443

Li Y, Ma D, Zhu M, Zeng Z, Wang Y. 2018. Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network. Accid Anal Prev. 111:354–63. doi:10.1016/j.aap.2017.11.028

Li X, Lord D, Zhang Y, Xie Y. 2008. Predicting motor vehicle crashes using support vector machine models. Accid Anal Prev. 40(4):1611–18. doi:10.1016/j.aap.2008.04.010

Madanat S, Liu P-C. 1995. A prototype system for real-time incident likelihood prediction. ITS-IDEA Program Project Final Report [accessed 2019 Jul 26]. https://trid.trb.org/view/465172.

McDonald AD, Lee JD, Schwarz C, Brown TL. 2018. A contextual and temporal algorithm for driver drowsiness detection. Acc Anal Prevent. 113(January):25–37. doi:10.1016/j.aap.2018.01.005

Michaels J, Chaumillon R, Nguyen-Tri D, Watanabe D, Hirsch P, Bellavance F, Giraudet G, Bernardin D, Faubert J. 2017. Driving simulator scenarios and measures to faithfully evaluate risky driving behavior: A comparative study of different driver age groups" ed. Jun Xu. Plos One. 12(10):e0185909. doi:10.1371/journal.pone.0185909

Mousannif H, Sabah H, Douiji Y, Sayad Y O. 2016. Big data projects: Just jump right in!. Int J Pervasive Comp & Comm. 12(2):260–288. doi:10.1108/IJPCC-04-2016-0023

Nguyen HM, Cooper EW, Kamei K. 2011. Borderline over-sampling for imbalanced data classification. International Journal of Knowledge Engineering and Soft Data Paradigms 3(1): 4. doi:10.1504/IJKESDP.2011.039875

Perez MA, Sudweeks JD, Sears E, Antin J, Lee S, Hankey JM, Dingus TA. 2017. Performance of basic kinematic thresholds in the identification of crash and near-crash events within naturalistic driving data. Accid Anal Prev. 103:10–19. doi:10.1016/j.aap.2017.03.005

Ragab A, Craye C, Kamel MS, Fakhri K. 2014. A visual-based driver distraction recognition and detection using random forest. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 8814: 256–265.

Siddiqui C, Abdel-Aty M, Huang H. 2012. Aggregate nonparametric safety analysis of traffic zones. Accid Anal Prev. 45:317–25. doi:10.1016/j.aap.2011.07.019

Shi X, Wong YD, Li MZF, Chai C. 2018. Key risk indicators for accident assessment conditioned on pre-crash vehicle trajectory. Accid Anal Prev. 117:346–56. doi:10.1016/j.aap.2018.05.007

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. 2014. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res. 15: 1929–58.

Theofilatos A, Chen C, Antoniou C. 2019. Comparing machine learning and deep learning methods for real-time crash prediction. Transport Res Record. 2673(8):169–178. doi:10.1177/0361198119841571

Vapnik VN. 1995. The nature of statistical learning theory. The nature of statistical learning theory. New York: Springer.

Wang L, Shi Q, Abdel-Aty M. 2015. Predicting crashes on expressway ramps with real-time traffic and weather data. Transport Res Record. 2514(1):32–38. doi:10.3141/2514-04

Werneke J, Vollrath M. 2013. How to present collision warnings at intersections? - A comparison of different approaches. Accid Anal Prev. 52:91–99. doi:10.1016/j.aap.2012.12.001

WHO. 2017. Road Traffic Injuries [accessed 2019 Jul 18]. https://www.who.int/en/news-room/fact-sheets/detail/road-traffic-injuries.

Yan X, Zhang Y, Ma L. 2015. The influence of in-vehicle speech warning timing on drivers' collision avoidance performance at signalized intersections. Transport Res Part C: Emerg Technol. 51:231–242. doi:10.1016/j.trc.2014.12.003

Yu R, Abdel-Aty M. 2013. Utilizing support vector machine in real-time crash risk evaluation. Acc Anal Prevent. 51:252–259. doi:10.1016/j.aap.2012.11.027

Yu R, Zheng Y, Abdel-Aty M, Gao Z. 2019. Exploring crash mechanisms with microscopic traffic flow variables: a hybrid approach with latent class logit and path analysis models. Acc Anal Prevent. 125(December 2018):70–78. doi:10.1016/j.aap.2019.01.022