Transportation Research Interdisciplinary Perspectives 31 (2025) 101457

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

journal homepage: www.sciencedirect.com/journal/transportation-research-interdisciplinary-perspectives

A stacked ensemble model for traffic conflict prediction using emerging sensor data

Bowen Cai [a,*], Léah Camarcat [a], Nicolette Formosa [b], Mohammed Quddus [a]

[a] Imperial College London, UK
[b] National Highways England, UK

ARTICLE INFO

Keywords:
Traffic Conflicts Prediction
Video Processing
Stacked Ensemble Learning
Recursive Feature Elimination

ABSTRACT

Over recent decades, a plethora of Safety Surrogate Measures (SSMs) have emerged as valuable metrics for predicting traffic conflicts. However, existing research predominantly focuses on conflict prediction at junctions or relies on a single SSM, such as time-to-collision, for detecting vehicle-related conflicts. This limitation highlights the challenge of selecting appropriate SSMs for vehicle- or segment-based conflict prediction, considering the diverse range of factors influencing conflict outcomes. To address this gap, this study leverages data collected from various infrastructure and vehicle-based sensors, including drones, lidars, radars and cameras, across different scenarios in China and the UK: urban junctions, motorway segments and vehicle-based data from instrumented vehicles. Employing machine learning approaches to handle the extensive and disaggregated data, a novel stacked ensemble learning model is proposed. This model integrates a Random Forest (RF), three-layer Deep Neural Networks (DNN), Support Vector Machine Radial (SVM-R), and a Gradient Boosting Model (GBM) meta layer to enhance prediction accuracy. The Recursive Feature Elimination (RFE) algorithm is then employed to identify the most influential SSMs for conflict prediction in each scenario. Results demonstrate the superiority of the stacked ensemble learning model, achieving accuracies of 88 % for junctions, 87.5 % for motorway segments and 99 % for vehicle-based conflicts. Furthermore, the study highlights the necessity of employing different SSMs for conflict detection in various scenarios. These findings hold significant implications for roadway operators and vehicle manufacturers, aiding in the development of strategies to detect infrastructure and vehicle-related traffic conflicts.

1. Introduction

Traffic collisions were responsible for 1.19 million deaths worldwide in 2021 (World Health Organization, 2023). In response to the challenge posed by the scarcity of collision data, the concept of traffic conflicts has emerged as a valuable tool for evaluating traffic safety. Theoretical frameworks, including extreme value theory and crash probability analysis, have been employed to validate the relationship between traffic conflicts and collisions. Safety surrogate measures (SSMs) offer a means to assess conflict risk, drawing upon data collected by sensors such as cameras, radars, and lidars installed on various platforms, including drones, roadside units, and vehicles. The extraction of pertinent information from raw data, including vehicle positioning, speed, and acceleration, necessitates sophisticated computer vision and data processing techniques, enabling the computation of diverse indicators essential for comprehensive safety evaluation.

Existing studies have assessed the performance of several risk metrics derived from longitudinal manoeuvres against four evaluation criteria: accuracy, timeliness, robustness and efficiency (Lu et al., 2021)]. They demonstrated that not a single measure performs well in predicting a traffic conflict reliably. For instance, they found that metrics performing better in prediction accuracy such as Time-To-Collision (TTC) or single-step Probabilistic Driving Risk Field (S-PDRF) performed weakly in timeliness. Indeed, the literature contains mixed conclusions on best performing SSMs based on the context. For instance, Tak et al. (2020)] found that deceleration-based metrics perform better for rear-end conflicts than Time-To-Collision (TTC) and distance-based metrics whereas He et al. (2018) found that modified TTC (MTTC) outperformed TTC and Deceleration Rate to Avoid a Crash (DRAC). A significant research gap in literature is the absence of studies evaluating SSMs performance in

* Corresponding author.
E-mail addresses: b.cai22@imperial.ac.uk (B. Cai), leah.camarcat18@imperial.ac.uk (L. Camarcat), nicolette.formosa@nationalhighways.co.uk (N. Formosa), m.quddus@imperial.ac.uk (M. Quddus).

https://doi.org/10.1016/j.trip.2025.101457
Received 28 August 2024; Received in revised form 11 March 2025; Accepted 3 May 2025
Available online 14 May 2025

different road environments and SSM selection justification for specific context-dependent applications (Arun et al., 2021). Moreover, studies that include vulnerable road users and that consider other road types than signalised intersections are lacking in the literature (Arun et al., 2021). The transferability of common model architectures for conflict prediction across different road scenarios has not been examined and is essential for industrial deployment of predictive models.

The primary objectives of this study are two-fold: (i) develop a new stacked ensemble learning model to predict conflict accurately and assess the transferability of a common model in different scenarios, and (ii) to conduct an exhaustive comparative analysis of various SSMs, evaluating their effectiveness in identifying conflicts across three diverse road environments: (a) urban intersection data, (b) infrastructure-based motorway data, and (c) vehicle-based motorway data. Studying data from both an infrastructure and a vehicle-based viewpoint is essential as the infrastructure-based data captures all the vehicles passing in one point in space whereas vehicle-based data captures the same surrounding vehicles alongside the ego-vehicle's movements. This results in the need for different analysis to be carried on and confirms the context dependence of conflict studies.

This study tests and compares the performance of different combinations of stacking frameworks and found that a combination of having Random Forest (RF), three-layer deep neural networks (DNN), and support vector machine (SVM) as the first layer, gradient boosting model (GBM) as the meta layer performs the best in traffic conflict prediction in real time. The integration of RF, DNN, and SVM as strong classifier extract conflict features, and GBM takes the predictions from individual models in the first layer to make more robust prediction on traffic conflict achieves good performance in conflict classification. Recursive Feature Elimination (RFE) is employed for the evaluation of feature contribution to the output of each model and a new aggregation method of the different rankings is proposed.

The rest of this paper is structured as follows: Section 2 presents the related works; Section 3 puts forth the methodology with the data acquisition and processing as well as algorithms employed; Section 4 presents the results; Section 5 discusses the performance and contributions of this study while Section 6 concludes this paper.

2. Related works

2.1. Safety Surrogate measure performance evaluation

Several studies have examined the performance of Safety Surrogate Measures. Lu et al. (2021) assessed the performance of six risk metrics for longitudinal maneuvers against four evaluation criteria: accuracy, timeliness, robustness, and efficiency. They demonstrated that no measure performs well in all performance dimensions. For instance, they found that metrics performing better in prediction accuracy such as Time-To-Collision (TTC) or single-step Probabilistic Driving Risk Field (S-PDRF) performed weakly in timeliness. Johnsson et al. (2021) explored a relative approach to validate the effectiveness of SSMs in a case study of six intersections in northern Europe and found that TTC indicator was in line with the ground truth of bicycle infrastructure. Wang et al. (2021) identified limitations and opportunities of SSM application in Connected and Automated Vehicles (CAV) including SSM for CAV trajectory optimization and SSM for individual vehicles and vehicle platoon. Similarly, Xie et al. (2018) also found that conventional SSM such as TTC may not be good indicator in CAV scenarios because TTC neglects the potential risk of car-following scenarios where the following vehicle's speed is slightly less than or equal to the leading vehicle's but the spacing between two vehicles is relatively small. Nadimi et al. (2021) thinks it is unclear which SSM indicators are able to reflect danger more precisely. Novel methods should be developed to make comparison by taking the advantage of statistical analyses of SSM datasets. Therefore, comparing different types of SSMs and identifying the most effective ones in specific scenarios is important to improve the precision and accuracy of conflict prediction and safety evaluations without crash data.

Studies often distinguished between urban and motorway road environments. Regarding motorways, different findings on the best performing SSM are present in the literature. Tak et al. (2020) found that deceleration-based metrics perform better for rear-end conflicts than TTC and distance-based metrics. The study evaluated three SSMs and performed a comparative analysis between collision risk and driver behaviour based on two different car-following theories: action point model and asymmetric driving behaviour model. He et al. (2018) found that MTTC outperformed TTC and DRAC using the statistical relationship between crash records and the safety measures. Regarding urban environments and more particularly intersections, few studies have comparatively evaluated the performance of SSM while accounting for different road users. Deceleration-based SSM have been proven to be more appropriate for unsignalized intersections from the literature (Bonela and Kadali, 2022). Simulation models have also been employed extensively to assess the performance of SSMs at intersections such as a study comparing the performance of the Aggregate Conflict Propensity Metric compared to other SSM for 12 signalised intersections (Wang and Stamatiadis, 2014). However, most of the studies are limited by the number of SSMs studied and compared as well the single road environment considered.

Many methods exist to assess feature importance for classification, notably for feature selection tasks. Dimensionality reduction is one of the most commonly employed methods but does not allow for individual comparison of each feature as these methods project features in a lower dimension feature space and new features are built as a combination of features. Feature selection methods can be categorized into supervised, semi-supervised and unsupervised classes depending on the availability of data. For the purpose of this paper, only supervised methods will be discussed. Supervised methods or flat features can be further subdivided in three classes which are filter, wrapper and embedded models (Tang et al., 2014). Filter and wrapper models are distinguished by the absence for the former or use of a learning algorithm to assess the feature importance for the latter. Similarly to wrapper models, embedded models employ a learning algorithm but assesses the feature importance during learning to increase computational efficiency. As this paper looks at features extensively used in predictive modelling, embedded models were selected for their efficiency.

2.2. Data-driven conflict prediction methods

Conflict modelling includes both conflict frequency and conflict severity models (Zheng et al., 2021). This paper will only focus on frequency with the aim of predicting conflicts in real-time as a binary classification problem. There exists a plethora of statistical and data-driven conflict prediction methods. The most common algorithms used for conflict prediction are Neural Networks (NN) with Long Short-Term Memory NN, Convolutional NN, Deep NN and Feed Forward NN, Support Vector Machine and Bayesian Networks (Marcillo et al., 2022). Random forest (RF) is a widely used strong classifier which combines multiple decision trees to obtain a classification result through a voting mechanism. Random forest outperforms decision trees by handling well overfitting problem in which each decision tree will not grow deep and different trees are trained to learn different features. RF is an improvement over bagged decision trees that combines predictions from multiple models in ensembles and performs the best if the predictions from the sub-models are uncorrelated or weakly correlated (Berk, 2016). RF changes the bagging algorithm with learned sub-trees resulting in less correlation in the subtrees' prediction. RF has been widely used in traffic conflict prediction for having strong classifying ability, identifying and ranking influencing factors affecting traffic conflicts or crashes, and extracting latent features of traffic data. Yan and Shen (2022) built random forest model with Bayesian optimization to predict traffic accidents and identify influencing factors affecting unsafe traffic conditions. Zhao et al. (2022) utilized deep convolutional neural network and random forest model to initiate a risk analytic model to predict traffic accident severity. Wang et al. (2021) built a random forest model to avoid congested road and choose high-speed route to avoid traffic congestion.

Gradient Boosting Machine (GBM) is an ensemble method sequentially combining the predictions of multiple weak learners and decision trees. Contrasting with random forests that rely on simple averaging of multiple models to make predictions, boosting methods use a different approach for forming ensembles (Natekin and Knoll, 2013). Boosting involves adding new models to the ensemble sequentially. At each iteration, a new weak base learner is trained to address the errors made by the ensemble up to that point (Natekin and Knoll, 2013). GBM builds models sequentially and the subsequent models reduce the errors of the previous model through minimizing loss function with optimal gradient descent in each iteration step. Similar to Random Forest, the basic prediction model used for GBM is decision trees (Hastie et al., 2001).

In gradient boosting machines (GBMs), the learning process involves sequentially fitting new models to improve the accuracy of the response variable estimate. GBM algorithm creates new base learners that are highly correlated with the negative gradient of the loss function, which reflects the errors of the entire ensemble (Natekin and Knoll, 2013). GBM combines multiple weak prediction models, often decision trees, in a sequential process to improve prediction accuracy (Natekin and Knoll, 2013; Hastie et al., 2001). Each subsequent model in the sequence focuses on correcting the errors made by its predecessor, resulting in a stronger, more robust model. GBM is particularly effective for predicting the likelihood of traffic conflicts by analyzing factors such as vehicle speed, distance, and lane changes (Zhang and Haghani, 2015; An et al., 2022; Niyogisubizo et al., 2023). Compared to simpler models, GBM offers greater precision in identifying potentially risky traffic situations (Zhang and Haghani, 2015; An et al., 2022; Niyogisubizo et al., 2023).

SVM (Support Vector Machine) is a supervised machine learning algorithm which is mainly used to classify data into different classes (Berk, 2016). Unlike most algorithms, SVM makes use of a hyperplane which acts like a decision boundary between the various classes (Berk, 2016). A main advantage of SVM is that it can be used for both classification and regression problems. In general, SVM is used for classification problem, and the SVR (Support Vector Regressor) is used for regression problems.

SVM can be used for classifying non-linear data by using the kernel trick. The kernel trick means transforming data into another dimension that has a clear dividing margin between classes of data (Berk, 2016). A kernel can be used to transform data into another dimension that has a clear dividing margin between classes of data. Kernel functions offer the user the option of transforming nonlinear spaces into linear ones. The kernel trick enables the identification of the optimal hyperplane between various classes of data (Hemsi, et al, 2020; Hofmann et al., 2008a, b). Therefore, complicated non-linear classification problem can be solved by finding a suitable hyperplane in a higher dimension by using the kernel trick. Orsini et al. (2021) used support vector machine as classifier, and Monte Carlo cross-validation for improving model robustness to make real time traffic conflict prediction within 5 min time duration and achieved 93 % accuracy. SVM can also be used to analyze causalities along with traffic conflict prediciton. Zhong and Du predict traffic casualties using SVM with heuristic algorithms based on collision data of urban roads in China (Zhong and Du, 2023).

A Deep Neural Network (DNN) is a type of artificial neural network characterized by having multiple layers between the input and output layers, including neurons, synapses, weights and biases, and activation functions. These layers and components enable DNNs to model complex relationships and patterns in data effectively. DNN is widely used in transportation and can be designed to predict potential collisions or near-misses on the road by analyzing real-time traffic data, including vehicle positions, speeds, and trajectories (Tang et al., 2019; Xiao, 2019). By leveraging its ability to identify complex patterns and relationships in data, a DNN can detect indicators of high-risk scenarios and provide early warnings (Formosa et al., 2020; Formosa et al., 2023). This capability enables proactive interventions to mitigate risks and enhance road safety, making it a valuable tool for real-time traffic management and collision prevention systems (Formosa et al., 2020; Formosa et al., 2023).

A single machine learning model can be used to classify conflicts, but wouldn't a combined structured model be more robust and powerful?

The stacked ensemble learning approach integrates predictions from several machine learning models into a single *meta*-learner model to make robust prediction and improve the overall model's generalizability (Kalule et al., 2023). The robustness achieved in stacked ensemble models are the accumulation of performances from several estimators enables them to capture the nonlinearities in the dataset (Kalule et al., 2023). However, only a few research applied stacked ensemble models to make traffic conflicts prediction. Some related studies are Bokaba (2022) used ensemble models for predicting traffic congestion. It compared prediction accuracy with different combinations of stacking models, and found that the combination of AdaBoost with decision trees exhibited the best performance (Bokaba et al., 2022). Ahmad (2023) used heterogeneous ensemble learning for enhanced crash forecasting. The stacking model applied by this study is the integration of decision tree, random forest, negative binomial, Poisson regression, and gradient boosting as the base-learner, random forest with grid search optimization as meta learner. The study showed that the performance of stacking model outperforms individual machine learning model as well as count data model (Ahmad et al., 2023). Tang (2019) put forwarded a two layer stacking framework with random forest, AdaBoost, and Gradient Boosting Decision Tree as the first layer. A simple Logistic Regression model was set as the second layer to make prediction on traffic crash. Results showed that stacking model achieves superiro performance with comparison to other typical machine learning models (Tang et al., 2019). Additionally, Xiao (2019) used support vector machine and KNN ensemble learning for traffic incident detection and found that stacking model achieved good performance in small likelihood incidents and maintain model robustness as well (Xiao, 2019). Zhang (2022) applied stacking model integrated with random forest, support vector machine, long short-term memory network, an attention mechanism as first layer model, and bidirectional LSTM as meta learner to make prediction for the recognition of the pedestrian crossing intention. It said that the proposed stacked model shows greater advantages in recognition accuracy with 95.36 % when the model is recognized at 0.5 s before crossing the zebra crossing (Zhang et al., 2022).

At the beginning of this study, 9 machine learning models were identified as adequate models for conflict prediction using classification methods:

- Logistic regression
- Naïve Bayes
- Support Vector Machine (linear kernel, polynomial kernel, gaussian kernel radial kernel)
- Decision tree
- Random Forest (RF)
- k-Nearest Neighbours (kNN)
- Gradient Boosting Method (GBM)
- Artificial Neural Network (ANN)
- Deep Neural Network (3 layers)

All combinations of models were tested as well as the number of models included in each combination. The combination with three models in the first layer and one model in the second layer performed the best. This can be explained by finding the tradeoff between overfitting and underfitting for the first layer where more models can capture more features, whereas the second layer benefits from having only one model to make the final decision. Firstly, the combination of one ensemble model (RF), one functional analysis model (SVM) and one deep learning model (DNN) allows the capture of different non-linearities in the data. Secondly, RF is more powerful than logistic regression, naïve bayes and decision tree because multiple decision trees are employed with a voting mechanism to handle overfitting. Support Vector Machine with a radial kernel was found to be the best performing model because of its efficiency for classifying non-linear data. Deep Neural Network is commonly employed for complex pattern recognition and classification and is a very accurate method to extract features. The second layer employs a GBM because this ensemble method is efficient at capturing residual patterns that could be present in the previous layer, and is robust to overfitting.

3. Methodology

3.1. Data acquisition and processing

The most challenging aspect in traffic conflict studies is to detect 'actual' traffic conflicts from data captured by sensors including drones, roadside camera, lidar, radar, etc. The next sections present dynamic traffic data collection methodology.

3.1.1. Infrastructure-based data acquisition

The infrastructure-based data was collected with an Unmanned Aerial Vehicle (UAV) flying over five typical junctions s in Shanghai and Suzhou, China for the junction data and motorway segments on the Sutong Bridge in China for the motorway data.

For junction scenes, data was collected on typical road scenarios, four representative junctions were picked: one-way street, waiting area junction, multi-directional junction, and ordinary junction. Drones were used to take aerial photography and record aerial video in the daytime with normal traffic conditions between 10am and 12 pm (off-peak hours) and 4 pm–6 pm (peak hours) respectively. For the motorway segments on the Sutong Bridge, data was collected between 11am and 2 pm. Four roadway segments were picked in dense area to record videos.

3.1.2. Vehicle-based motorway data collection

The vehicle-based data was collected in the UK on the M1 motorway with a RGB monocular camera installed in an instrumented vehicle alongside a GPS. The vehicle was driven on a 55 km stretch between junctions 17 and 23 of the M1 motorway with 5 different drivers at varying times between 10am and 7 pm to encompass a wide variety of traffic scenarios and weather conditions.

3.1.3. Computer vision and feature extraction

The data was first processed using computer vision with YOLO (You Only Look Once) v8 to detect road users and identify the category: vehicle, HGV, cyclists or pedestrian. From the bounding box, the real-world position was obtained through spatial coordinates transformation from the image space to the physical space. For infrastructure-base datasets, the vehicles' length and width were also computed using the bounding boxes. Yolo-v8 has built-in multi-object tracking model and was employed to track the detections from frame to frame. Calibration was needed to account for noisy object detection where the bounding box moved slightly from one frame to another. This was performed by smoothing the positions over several frames and by fine-tuning the detection models to increase the accuracy and the precision of the detections. From the real-world positions and the time elapsed between each frame, the velocity and acceleration could be calculated which enabled the calculation of all selected SSM between two road users. For the vehicle-based motorway scenario, only the data extracted from the closest vehicle to the ego-vehicle is included in the final dataset.

The detection logic with application of YOLOv8 is explained in the following figure.

The following flow chart exhibits the whole procedures including detection, tracking, smoothing, and parameter calculating to get the needed data from drones.

The following figures depicts the recognition and detection of dynamic objects on the motorway.

3.1.4. Conflict identification

Traffic conflicts were identified through human judgement for both infrastructure-based motorway and urban junction datasets collected by drone detection (Federal Highway Administration (FHWA), 1989). However, experts were trained to identify traffic conflicts with the identification of evasive maneuvers such as braking, swerving and acceleration. During the training phase, the conflicts were cross validated by verifying that the associated TTC values were inferior to 3 s. The TTC threshold used in previous studies has varied from 2 to 4 s (Berk, 2016; Yan and Shen, 2022). In order to include a greater number of conflict samples in this research, TTC* was set as 3 s at junctions scenarios for the UAV data (Zhao et al., 2022). The reason why TTC was not used directly as a judgment standard is due to the potential correlation between conflicts and conflict affecting factors. If TTC was used directly to identify conflicts and TTC, MTTC, and other SSMs were used as independent variables to predict conflicts, the prediction accuracy would be very high but would have been meaningless because of the strong correlation is existed between the independent variables and the dependent variable.

Human observers' perception of traffic conflicts is an appropriate way to determine ground truth at junction scenarios (Wang et al., 2021a, b). Human observers are recruited from experts of transportation, identifying evasive maneuvers on roadways. Many studies have proved this and suggest that human judgements are reliable and valid for detecting severity in traffic events (Natekin and Knoll, 2001). Hydén (1987) showed that with the right initial training, there is good agreement reached between observers in terms of identifying traffic conflicts (Hydén, 1987). Van der Horst (1984) also showed that humans have an innate ability to assess the level of danger in a traffic situation especially when it is based on an actual observation (Van der Horst, 1984).

For UAV data, an initial screening of video data was performed to identify potential traffic conflicts using continuously calculated TTC for all collected videos. A TTC value inferior to 5 s triggered a manual check from a trained expert to determine whether the scene is a traffic conflict or not. A binary value (0, 1) was then assigned to every scene. The five second TTC threshold was selected conservatively because the majority of studies employed TTC thresholds comprised between 2 and 4 s (Mahmud et al., 2017, Li et al., 2014). The trained experts used the presence of an evasive maneuver such as braking, swerving or accelerating to identify a conflict. Many studies have proved the validity of conflict identification by experts and suggest that human judgements are reliable and valid for detecting severity in traffic events (Diwakar et al., 2023).

For vehicle-based data, this study employs the method proposed by Formosa et al. (2020) to identify traffic conflicts. This automatic identification method relies on computer vision to detect several indicators found in the literature. Rear-end conflicts were identified when specific conditions—such as lateral distance below 1.75 m, time headway under 3 s, and speed/braking dynamics between subject and leading vehicles—were met, alongside a braking threshold of ± 2.943 m/s^2 (Naito et al., 2009; Angkititrakul et al., 2011). Similarly, conditions required for lane change conflicts included total time headway between two vehicles inferior to 3 s, the lateral distance is comprised between 1.5 m and 2.2 m, and the longitudinal available space is less than 75 mm. Formosa et al. (2020) validated the method through manual video reviews (Saunier et al., 2010) (See Fig. 1, Fig. 2, Fig. 3).

3.2. Selected SSM

Thirteen commonly used SSMs were selected from total of 28 SSMs in the literature review due to lack of data availability in real scenarios such as mass of vehicles, the released energy that may affect vehicle

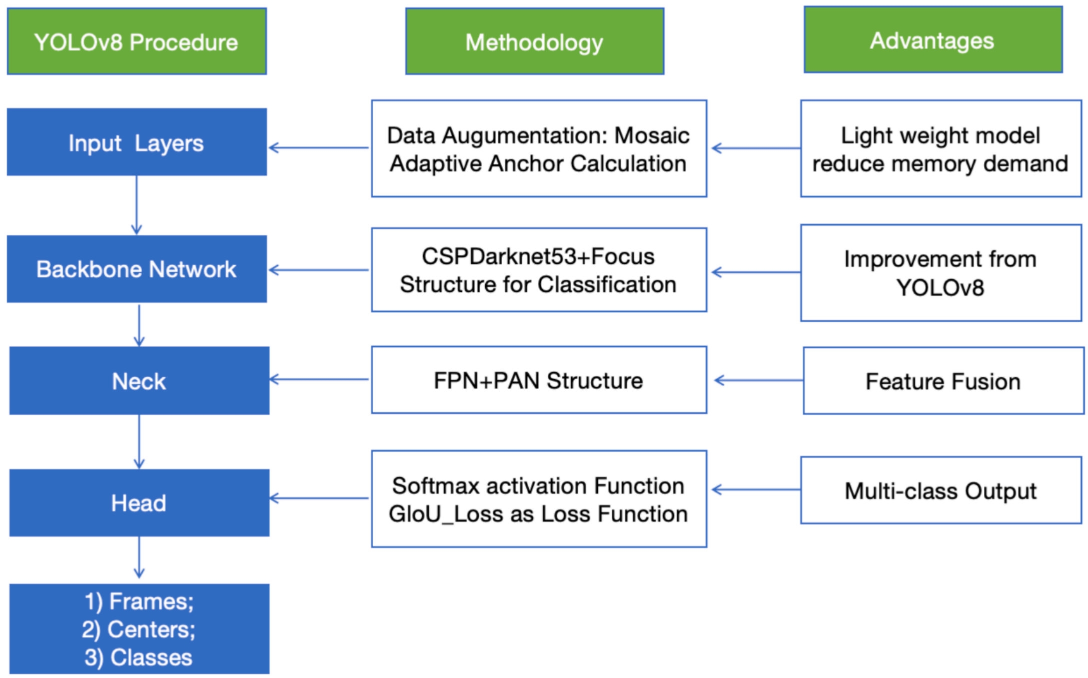

Fig. 1. Detection logic of drones.

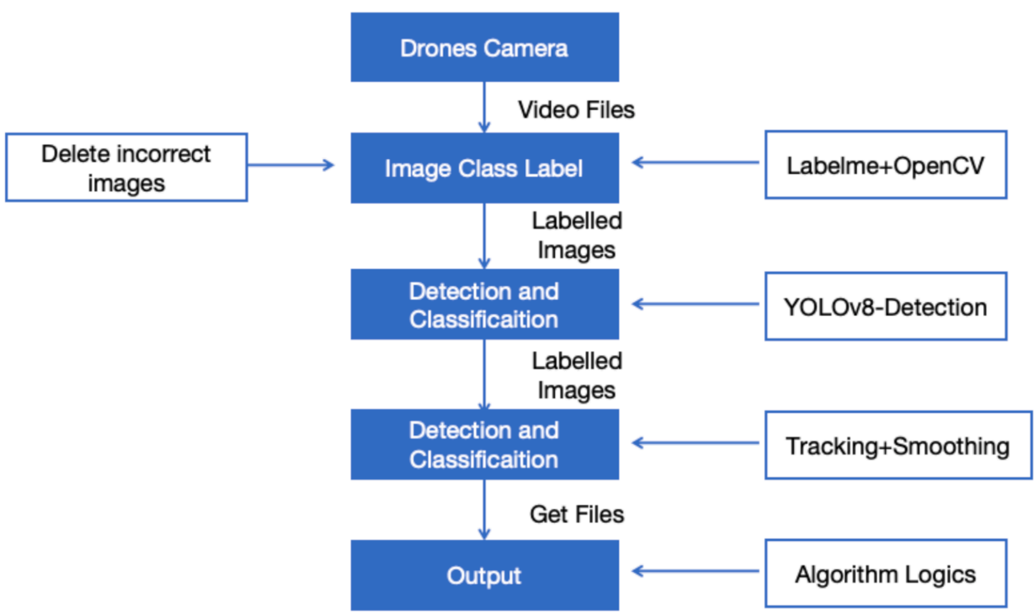

Fig. 2. Method overview of data extraction from drones.

occupants, and the change in total kinetic energy before and after the crash, etc. The selected SSMs involve time-based, distance-based and deceleration-based measuring metrics which are presented in Table 1. The time-based metrics include Time-To-Collision (TTC), modified TTC (MTTC), Post-Encroachment Time (PET), Time integrated TTC (TIT), Time Exposed TTC (TET), Headway (H), and Gap Time (GT). Among the selected distance-based metrics are headway, Proportional Stopping Distance (PSD) and Difference of Stopping distance and Safety distance (DSS). Deceleration-based metrics employed are Deceleration Rate to Avoid a Crash (DRAC), Potential Index for Collision with Urgent Deceleration (PIUCD) and Crash Index (CAI). Every SSM is calculated between two vehicles at a specific instant in time.

With D corresponding to the distance between two vehicles, ΔV and Δa the difference in speed and acceleration between two vehicles, v_l and v_f the velocity of the leading and following vehicles respectively, a_l and a_f the acceleration of the leading and following vehicles respectively, Δt the time step, TTC^* the threshold value of TTC, $MADR$ the maximum deceleration rate, μ is the friction coefficient and g is the gravity acceleration constant.

It is important to note that for the vehicle-based motorway scenario, only 6 SSMs were computed: TTC, MTTC, PET, PSD, DRAC, CIF. These six measures were selected to be representative of the time-based, distance-based and deceleration-based SSMs.

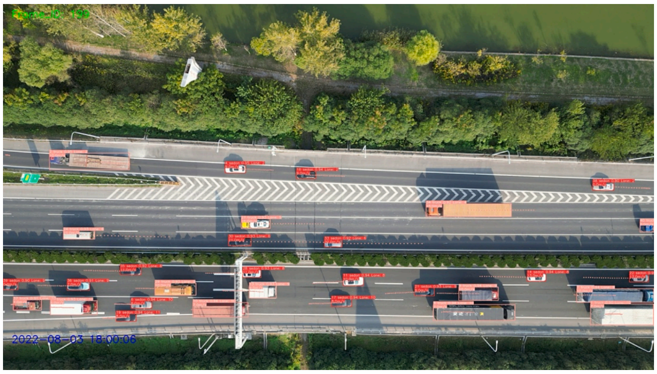

Fig. 3. Collecting traffic dynamics data on a motorway segment using a drone.

3.3. Stacked ensemble learning model for traffic conflict prediction

The predictive model is a stacked ensemble learning model composed of two layers and presented in Equation 15 and Fig. 4. In the first layer, three models run in parallel: (i) Random Forest (RF), (ii) Support Vector Machine Radial (SVM-R), and (iii) deep neural network comprising of three hidden layers. The results generated by these three models serve as inputs to the second layer which is a Gradient Boosting Machine (GBM) model whose output is the final binary prediction detecting the presence or absence of a conflict. Fig. 5, Fig. 6

$$y_{pred} = GBM(RF(X), SVMR(X), DNN(X)) \quad (15)$$

These four models and this architecture were selected for their highest performance among 60 combinations of 25 machine learning and ensemble models tested. Each model in the first layer is independently trained on the dataset. The models' outputs create a new training set to train the GBM forming the second layer. Various hyper-parameter fine-tuning techniques were applied across different models. The Adam optimizer was utilized for the DNN, while the Grid Search method was implemented for SVM, RF, and GBM. To mitigate overfitting, an early

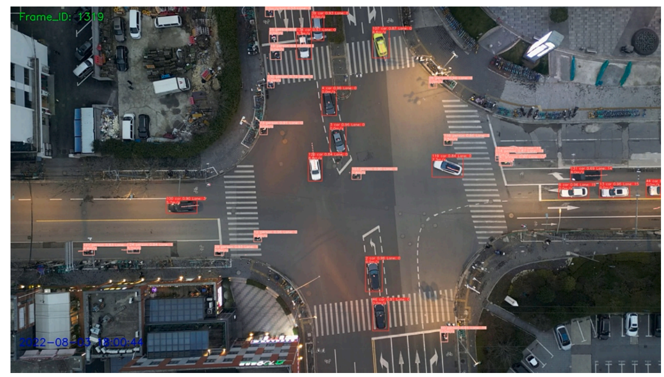

Fig. 4. Collecting traffic conditions data at a junction (using a drone).

Table 1
Definition of the selected safety surrogate measure.

Selected SSM	Definition	Equation
Time-To-Collision (TTC)	The time left to a collision assuming constant speed and direction.	$TTC(t) = \frac{D}{\Delta V}(1)$
Modified TTC (Ozbay et al., 2008)	Modified to consider acceleration discrepancies to account for all potential longitudinal scenarios.	$MTTC(t) = \frac{-\Delta V \pm \sqrt{\Delta V^2 + 2\Delta a D}}{\Delta a}(2)$
Post-Encroachment Time (PET) (Allen et al., 1978)	The period between the leading vehicle departure and the following vehicle arrival to a conflict area.	$PET = t_4 - t_2(3)$ t_2 is arriving time at a conflict point of the 2nd vehicle and t_1 is the time of the 1st vehicle departing
Time integrated TTC (TIT) (Minderhoud and Bovy, 2001)	TTC integrated over the period under a TTC threshold.	$TIT(t) = \sum_{n=1}^{N}\left[\frac{1}{TTC_{it}} - \frac{1}{TTC^*}\right]\Delta t(4)$ $TIT^* = \sum_{n=1}^{N} TIT(t)(5)$
Time Exposed TTC (TET) (Minderhoud and Bovy, 2001)	Summation of every instant where the TTC is below the threshold value.	$\#(6)\ TET(t) = \sum_{n=1}^{N}\delta_t \Delta t, \delta_t = \begin{cases} 1, 0 < TTC(t) < TTC^* \\ 0, otherwise \end{cases}$ (6) $TET^* = \sum_{n=1}^{N} TET(t)(7)$
Headway (Vogel, 2003)	The time between the front of the lead vehicle passing a point on the roadway and the front of the following vehicle passing the same point.	$H(t) = t_2 - t_1(8)$ t_2 is the arrival time of leading vehicle to a point on the roadway; t_1 is the arrival time of following vehicle to the same point.
Gap time	The time between the arrival of two vehicles to the conflict location.	$G(t) = t_3 - t_2(9)$ Where t_3 is the expected arrival at the potential point of collision, provided that the vehicle maintains the original approach speed and direction. t_2 is the time from the end of encroachment.
Proportion of stopping distance (PSD) (Astarita et al., 2012)	Proportion of the remaining distance to a collision point and the minimum acceptable stopping distance.	$PSD(t) = \frac{D}{\frac{v_f^2}{2MADR}}(10)$
Difference of stopping distance and space distance (DSS) (Okamura et al., 2011)	Difference between the distance between vehicles and the stopping distance.	$DSS(t) = \left(\frac{v_l^2}{2\mu g} + D\right) - \left(v_f^2 \Delta t + \frac{v_f^2}{2\mu g}\right)(11)$
Deceleration rate to avoid a crash (DRAC) (Saccomanno et al., 2008)	Deceleration rate needed to avoid a collision with the ratio between the difference in speed and their closing time	$DRAC(t) = \frac{\Delta V^2}{2D}(12)$
Potential Index for Collision with Urgent Deceleration (PIUCD) (Uno et al., 2002)	Distance separating two vehicles if they have completely stopped with emergency braking.	$PIUCD = \frac{v_l^2 - v_f^2}{2MADR} + D - v_f \Delta t(13)$
Crash index (CAI) (Ozbay et al., 2008)	Impact of speed on a potential collision's kinetic energy	$CAI = \frac{(v_f + a_f MTTC)^2 - (v_l + a_l MTTC)^2}{2} * \frac{1}{MTTC}(14)$

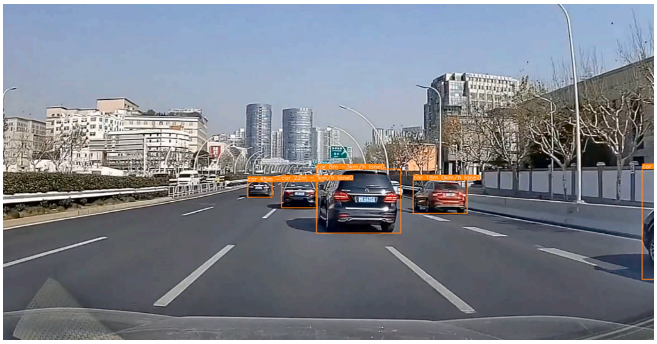

Fig. 5. Collecting vehicle-based data using an instrumented vehicle.

stop mechanism was employed. Furthermore, the dataset was divided into training and test sets in a 7:3 ratio.

The following figure presents the architecture of the proposed stacked ensemble learning model.

The stacked ensemble learning approach integrates predictions from several machine learning models into a single *meta*-learner model to make robust prediction and improve the overall model's generalizability (Kalule et al., 2023). The robustness achieved in stacked ensemble models results from leveraging the strengths of different machine learning algorithm as each individual model may excel at capturing different aspects of the dataset (Kalule et al., 2023). The next section introduces each individual model employed.

3.3.1. Random forest

Random forest is a classical and widely used strong classifier which combines multiple decision trees to get classification result through voting mechanism. Random forest outperforms decision trees by handling well with overfitting problem in which each decision tree will not grow deep and different trees are trained to learn different features. Random forest acts as the advanced version of bagging method to enhance the performance and predictive power of the ensembled decision trees. Random forest is also used for model interpretability by analyzing the contributing importance of each independent variable (Hastie et al., 2001).

The predicted value for random forest is:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{22}$$

Where B is bagging times, $f_b(x')$ is the prediction of individual regression trees on the sample of x' in regression problem. For classification problem, it randomly selects observations, builds a decision tree, and takes the average result through majority votes.

In the proposed stacked ensemble model, hyperprameters used by Random Forest are as follows:

number of trees: 500 The quantity of decision trees in the forest.

Number of variables randomly sampled as candidates at each split: 5.

Max_depth: 5 Maximum depth of each decision tree.

Min_sample_split:3 Minimum samples required to split a node.

Min_samples_leaf:5 Minimum samples required in a leaf node.

Max_sample:0.2 Fraction of the original dataset is given to any individual tree.

Random forest can be parallelized, making it scalable to large, high-dimensional datasets, which is just the case of the extensive SSM data. The interaction between random forest model with high-dimensional SSM data are explained in details as follows (Beest, et al., 2017):

- Non-linearity: SSMs often capture complex, non-linear relationships between variables. Random Forest can model these non-linear interactions effectively by capturing different aspects of the data.
- Robustness to Noise: SSM data can be noisy due to measurement errors or variability. Random Forest's ensemble approach averages out the noise across multiple trees, leading to more reliable predictions.
- Interpretabiltiy: Feature importance scores returned by Random Forest can be valuable in understanding which SSMs are more influential in predicting safety outcomes.
- Overfitting: High-dimensional SSM data is prone to overfitting during model training. By randomly selecting subsets of features, the model introduces variability, preventing it from becoming overly dependent on any single feature and thereby reducing the risk of overfitting.

Therefore, Random Forest should be suited for handling high-dimensional Safety Surrogate Measures data due to its inherent feature selection, robustness to noise, and the ability to model complex, non-linear relationships. Its ensemble nature and bagging technique make it resistant to overfitting, while its scalability and interpretability make it practical in explaining models built with high-dimensional SSMs data.

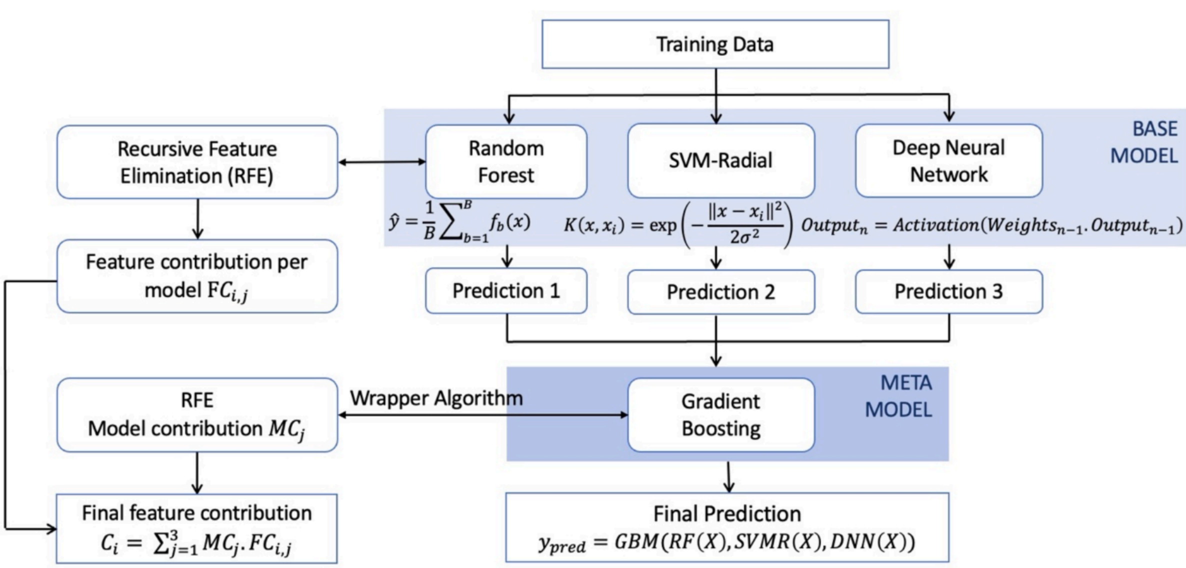

Fig. 6. Stacked ensemble learning model and variable contribution assessment architecture.

3.3.2. Radial Support Vector Machine (SVM)

Radial Support Vector Machine (SVM) is an algorithm established for both classification and regression. SVMs have different built-in kernel methods such as Gaussian, Linear, Radial, and Polynomial, which solve complex optimization and classification problems by applying linear estimation methods on a higher dimensional feature space through functional analysis (HofmDNN et al., 2008; Thurnhofer-Hemsi et al., 2020a,b).

The kernel function for SVM is the radial basis function with expression as:

$$K(x,x_i) = \exp\left(-\frac{||x-x_i||^2}{2\sigma^2}\right) \quad (16)$$

It is the multivariate Gaussian funciton, where xi represents center parameter, σ represents smoothing parameter, x is the input variable vector.

SVM searches the optimal hyper-plane that has a maximum margin between the nearest positive and negative samples with the expression as:

$$\text{argmin}\frac{1}{2}||w||^2, s.t.y_i\left(w^T\phi(x)+b\right) \geq 1 \quad (17)$$

Where $\phi(.)$ is the feature-space transformation function, and

b is the linear classification bias.

The proposed SVM model uses GridSearch Cross Validation.

The best model was: SVC(C = 1, gamma = 1).

The best parameter values were: {'C': 1, 'gamma': 1, 'kernel': 'rbf'}.

Verbose = 1 combinations of parameters and folds.

SVMs is a popular tool for analyzing temporal data with a timestamp for each record. SVMs are well-suited to temporal data because they can handle large, high-dimensional data sets, and are effective for classification, regression, and time series prediction (Thissen et al., 2003).

- SVMs use a risk function that combines empirical error and a regularized term.
- SVMs can handle high dimensional data well, even with relatively small training samples.
- SVMs can exhibit good generalization ability for complex models.

SVM uses feature engineering to convert time series data into a format involving lagged variables, rolling statistics, and seasonal indicators (Tajiri et al., 2010). Specifically, lagged variables (e.g., speed at time t − 1, t − 2) can be created to capture past behavior. Rolling statistics (e.g., average speed over the last few seconds) can be computed to summarize temporal trends.

In addition, SVM's regularization parameter helps prevent overfitting, which is useful when dealing with high-dimensional temporal features. SVM does not inherently model sequential data. It relies on feature engineering to incorporate temporal information, which can be labor-intensive and may not capture long-term dependencies effectively, but is good at short term time series analysis (Thissen et al., 2003)..

3.3.3. Deep Neural Networks (DNN)

Deep Neural Networks (DNN) are built from units called perceptrons (ptrons). Ptrons are composed of one or more inputs, an activation function and an output. DNN combines ptrons in structured and connected layers, and the ptrons are independent of each other in the given layer [30]. The DNN used in the study has three hidden layers, with 5, 3, 3 ptrons respectively. Feedforward and resilient back propagation algorithm is used to train and update the weights.

Feed-word function:

$$Output_n = Activation(Weights_{n-1} \cdot Output_{n-1}) \quad (18)$$

Back propagation:

$$E_{n-1} = W_n^T \cdot E_n \quad (19)$$

Weight updating:

$$W_n = W_n \cdot L_r \cdot \frac{\delta E_{n+1}}{\delta W_n} \quad (20)$$

In the equations, L_r represents the learning rate, which is 0.1 in our study. E is our error function and W represents the weights. As the learning iteration (epoch) goes up, the weight is updated according to the slope of the derivative of the error function.

Hyperparameters used in the proposed DNN are as follows:

Momentum: 0.7.

Number of epochs: 30.

Batch size: 32.

Learning rate: 0.05.

Batch size: 32.

Grid Search is used in DNN.

DNN is suited for capturing temporal dependencies in traffic data by modelling how a sequence of events leads to a conflict. Unlike traditional methods, DNNs do not need to rely solely on lagged observations for learning (Fawaz et al., 2018). Instead, DNNs can learn a meaningful representation from a large input sequence, focusing on the aspects most relevant to the prediction task (Fawaz et al., 2018).

In order to avoid overfiting issue, we managed to make DNNs effective through:

1) using simpler DNN architectures with fewer parameters to avoid overcomplexity and better suit the smaller dataset size.
2) using regularization techniques such as L1/L2 regularization to penalize large weighs in the network
3) using dropout technique to randomly drop neurons during training and using early stopping technique when validation performance stops improving.

This study applies the above techniques to reduce overfitting and make DNNs effective.

3.3.4. Gradient boosting Machine (GBM)

Gradient Boosting Machine (GBM) is an ensemble method sequentially combining the predictions of multiple weak learners and decision trees. GBM aims to improve overall predictive performance by optimizing the model's weights based on the errors of previous iterations, gradually reducing prediction errors and enhancing the model's accuracy based on Loss function minimization in classification problem (Mason et al., 1999). The loss function of GBM for classification is as follows:

$$\text{Loss Function} = -\sum_{i=1}^{n} y_i \log(p) + (1-p)\log(1-p) \quad (21)$$

In this function, y_i is the observed values and p is the predicted probability.

For loss function of GBM for regression, mean squared error is used.

GBM builds models sequentially and the subsequent models reduce the errors of the previous model through minimizing loss function with optimal gradient descent in each iteration step. The prediction method used for GBM is decision trees (Hastie et al., 2009).

Hyperparameters used in the proposed GBM model is as follows:

min_samples_split = 500.

min_samples_leaf = 50.

max_depth = 8.

max_features = 'sqrt' (square root).

max_features = 7.

subsample = 0.8.

learning_rate = 0.05.

GBM deals with high-dimensional data and excels at dealing with multicolinearity existed in dataset. GBM is also good at model interpretability. In the stacked ensemble model, GBM is set to be the second layer. It takes all outputs from the first layer and analyze the contributing variable important scores of each factor.

3.3.5. Training procedure and hyper-parameter finetuning

Each model in the first layer is trained independently on the dataset and the resulting outputs create a new training set used to train the gradient boosting model in the second layer, with the inputs being the prediction generated by the based models. Moreover, different hyper-parameter fine tuning methods were used for different models. Adam optimizer is used for DNN. Grid search is used for SVM, random forest, and gradient boosting method. Early stop mechanism is applied to avoid overfitting. Train set, validation set, and test set are split according to the ratio 6:2:2.

3.4. Identifying important features

Feature importance evaluation has been performed throughout the stacked model using the Recursive Feature Elimination (RFE) method. Recursive feature elimination is a process where predictors are systematically removed from a full set based on their importance scores, which are calculated through cross-validation [33]. Applying RFE on the three models in the first layer outputs feature contribution $FC_{i,j}$ for each feature i and model j. Secondly, each model contribution MC_j is computed by applying RFE on GBM model in the second layer as the input features are the prediction from each model of the first layer. The final contribution C_i of each feature i is computed with a weighted sum as presented in Equation 22.

$$C_i = \sum_{j=1}^{3} MC_j.FC_{i,j} \tag{22}$$

3.5. Detection accuracy and post-processing for computer vision methods

Detection accuracy of computer vision methods is influenced by the quality of input images. To clarify, YOLOv8 was primarily used for processing drone-based datasets in this study. The clearness of the recorded image often decreases alongside the increase in flight altitude of the UAV. The height is selected to capture the entire junction including zebra crossing and ensure that objects are sufficiently large to be detected. Another important parameter for height selection is to allow that the space is sufficient to calculate the speed of the detected objects. Each junction and motorway segment had drone height varying from 40 m to 100 m. The following Fig. 7 and Table 2 depicts the relationship between drone detection accuracy with flying height. Fig. 7 shows the field of vision of the UAV mounted camera at different heights, while Table 2 presents detection accuracy for a range of heights. In the study, different model performances in both detection and tracking procedures are also compared to decide the best model in object detection and tracking. Yolo-v8 outperforms others, which is reflected in Table 3 Comparison of detection algorithms and Table 4 Comparison of tracking algorithms.(Fig. 8, Fig. 9, Fig. 10).

Fig. 7. Drone detection accuracy in tested scenarios.

Table 2
Detection accuracy in different scenarios.

Height	Accuracy	car	Two-wheelers	person
40 *m*	Detection	99.6 %	98.8 %	97.3 %
	Classification	98.9 %	98.7 %	96.6 %
	Tracking	98.6 %	98.4 %	96.5 %
	Speed	96.2 %	94.3 %	\
70 *m*	Detection	99.1 %	98.8 %	96.5 %
	Classification	98.7 %	98.4 %	96.2 %
	Tracking	98.0 %	97.7 %	95.8 %
	Speed	94.8 %	92.8 %	\
100 *m*	Detection	98.6 %	96.8 %	95.3 %
	Classification	98.3 %	96.3 %	94.6 %
	Tracking	97.6 %	96.0 %	94.2 %
	Speed	93.6 %	91.2 %	\

This study also compares different model performances in both detection and tracking procedures.

In the majority of cases, detection accuracy is better for low drone height. If the drone flies too high (> 150 m), the object on the ground is too small to be detected, resulting in compromised detection accuracy. Post-processing smoothing using the moving average technique and re-identification methods were also applied to YOLO-v8. Post-processing methods aims to smooth the tracking ID, particularly reducing the loss of object ID due to coverage and interference from trees, buildings, big trucks or buses, and ID jumps. These issues are often caused by occlusions such as vehicle and pedestrian passing through.

4. Findings and discussion

Table 2 presents the outcomes for both the stacked ensemble learning model and Table 3 presents the result of the ablation study. The vehicle-based motorway scenario achieved the highest overall accuracy of 99.3 % with a sensitivity rate of 99 % for conflict detection. The infrastructure-based motorway and urban junctions scenarios also demonstrated relatively high accuracy, reaching 87.5 % and 88 % respectively, with a sensitivity rate of 87.9 % and 89.6 % for correctly classified conflicts. Vehicle-based scenarios demonstrate high classification accuracy. However, predicting conflicts using infrastructure-based sensors poses slightly more difficulty. This is attributed to the higher number of conflicts present in the infrastructure-based data and the higher heterogeneity in driving behaviour captured by the drones compared to the 5 different drivers collecting the vehicle-based data. Another potential factor is that conflicts defined from a human

Table 3
Comparison of detection algorithms.

Model	Algorithms	Pros and Cons
YOLO-v8	1) YOLO does not show the process of obtaining region proposals; 2) YOLO is unified as a regression problem	High Accuracy, Fast Speed
U-Net	Used for semantic segmentation, not for detection purpose	Slow Speed
Faster-RCNN	Although RPN and Fast R-CNN share convolutional layers in Faster R-CNN, it is necessary to repeatedly train the RPN network and Fast R-CNN network during model training.	High Accuracy, Low Speed
SSD	Unlike YOLO which uses a fully connected layer at the end, SSD directly uses convolution to extract detection results from different feature maps.	Fast Speed, Accuracy not good

Table 4
Comparison of tracking algorithms.

Model	Algorithm Logic	Pros. And Cons.
YOLO-v8	One stage model, combine detection and re-ID into the same process.	Fast Speed, High Accuracy
Fairmot	Multiple tracks execute Kalman filtering in parallel, which is faster than deepsort's Kalman filtering. Only one feature vector is saved, and exponential sliding average is used to fuse multiple feature vectors to speed up calculation. Anchor Free, accurate positioning	High Accuracy, Medium Speed
Deepsort	Two stage network, detection and re-ID are separated	High Accuracy, Low Speed
SDE	Detection and Embedding are divided into two independent models, with high performance but poor real-time performance.	High Accuracy but Low Speed
Kalman Filtering	Conventional tracking algorithm, not very useful in deep learning	Low Accuracy, Fast Speed

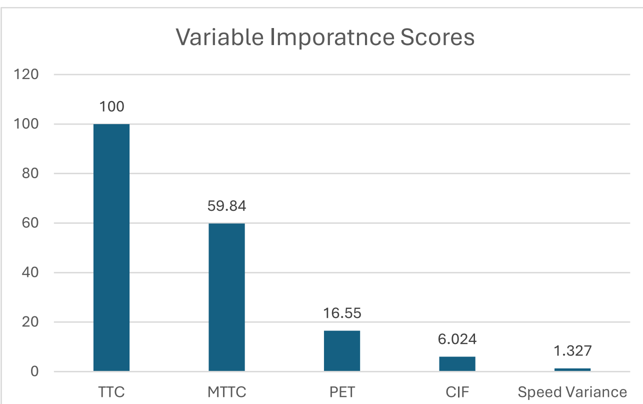

Fig. 8. Visualization of influential features in vehicle-based motorway scenario.

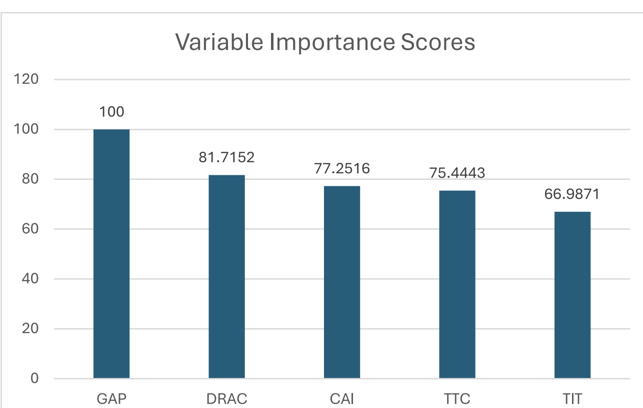

Fig. 9. Visualization of influential features in infrastructure-based motorway scenario.

perspective may not always equate to actual conflicts, and the prediction model struggles to adequately account for tolerance. Sensitivity and specificity precision closely mirror the accuracy of predictions, indicating minimal disparity in false positive and false negative ratios. (See Table 5.)

Analysis of ablation study revealed that, Gap Time, DRAC and CAI emerged as the most effective measures both for infrastructure-based urban junctions and motorway segments data. However, disparities

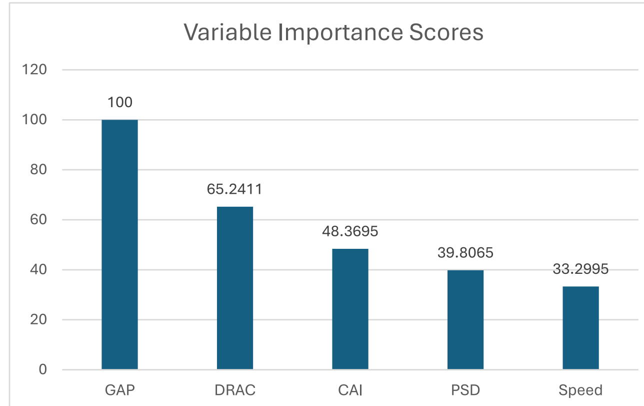

Fig. 10. Visualization of influential features in urban junction scenario.

were observed between motorway segments (infrastructure) and vehicle-based data: while time-based measures exhibited superiority for the latter, a mix of different measure types was evident for the former. Conversely, the primary influencing factors for vehicle-based conflict detection differed from those in infrastructure-based urban and motorway contexts. Specifically, deceleration rate emerged as more critical in infrastructure-based urban and motorway scenarios, whereas vehicle-based conflicts were more closely associated with time-based SSMs. The reason that Gap time, DRAC, CAI are more influential for infrastructure-based conflicts while TTC, MTTC, PET are more influential for vehicle-based conflicts is possibly due to their belongings of different classes of SSM. TTC, MTTC, PET, Gap time are time based SSM, but DRAC belongs to deceleration based SSM and CAI belongs to energy-based SSM (Wang et al., 2021). For vehicle-based conflicts, time based SSM are major safety indicators to measure the time proximity to a crash occurrence. However, for infrastructure-based conflicts, various SSM should be involved to measure different aspects of a crash occurrence. It also indicates that causes of infrastructure-based conflicts are more complicated than vehicle-based conflicts since more interactions (vehicle to vehicle and vehicle to person) occurred in infrastructure based scenarios than vehicle based scenarios (ego vehicle video collection). It can be also be ascribed to external drones capture all roadway conditions in all lanes contrast with just using single-lane following scenario for ego vehicle video. Ego vehicle video recording may ignore the surrounding situations so that energy based metrics are not relevant.

In addition, the significant contribution of CAI indicates higher variability in crash energy for larger traffic flows especially in urban junctions and dense motorways since congested traffic roadways are the places where interactions are more likely to happen. Frequent lane changes, vehicle merging, the conflict between vehicle passing and pedestrian crossing in the intersections, emergency braking due to less car following distance, frequent acceleration and deceleration, and multi-directional flows in urban junctions and motorways all increase the likelihood of collision occurrences especially for rear-end collision and vehicle merging collisions. Therefore, CAI becomes a good indicator to measure traffic characteristics in infrastructure based scenarios as more complicated behaviors (merging, crossing, lane changing) are taking place.

Furthermore, there are more sidesweep conflicts in intersections which could be more difficult to detect using only one type of SSM, for example just time based SSMs. (See Table 6.)

Variable correlations are also investigated. As shown in Fig. 11, significantly positively correlated features are Speed and Flow, Occupancy and Density, Density and Flow. It shows that as speed increases, traffic volume may also increase. As traffic volume increases, density increases, followed by the increase of occupancy. Significantly negatively correlated features are Flow and Headway, Occupancy and Average Speed, Speed and Density, Speed and Occupancy. When traffic

Table 5
Results of the stacked ensemble learning model for each scenario.

Scenario	Accuracy	Sensitivity	Specificity	F1 Score	Pos Pred Value	Neg Pred Value
Urban junctions – infrastructure-based	0.88	0.896	0.865	0.8823	0.8600	0.9000
Motorway segments – infrastructure based	0.875	0.879	0.871	0.8765	0.8700	0.8800
Vehicle-based motorways	0.993	0.99	0.986	0.9932	0.9884	1

Table 6
Results of the feature contribution assessment for each scenario.

Scenario		RF	SVM	DNN	Five most significant variables
Urban junctions – infrastructure-based	Model contribution	87 %	12 %	1 %	
	Most significant variables per model	GAP, DRAC, CAI, PSD, Speed	GAP, CAI, TIT, TTC, MTT	DRAC, Speed, MTTC, TIT, PIUCD	GAP, DRAC, CAI, PSD, Speed
Motorway segments – infrastructure based	Model contribution	84 %	13 %	3 %	
	Most significant variables	GAP, DRAC, CAI, TTC, TIT	GAP, CAI, TTC, TIT, MTTC	CAI, DRAC, DSS, GAP, Distance	GAP, DRAC, CAI, TTC, TIT
Motorway segments – Vehicle-based	Model contribution	70 %	26 %	4 %	
	Most significant variables	TTC, MTTC, PET, CIF, Speed Variances	TTC, MTTC, PET, CIF, Speed Variances	CIF, SpeedCar, Speed Relative variance, SD Density, Headway	TTC, MTTC, PET, CIF, Speed Variances

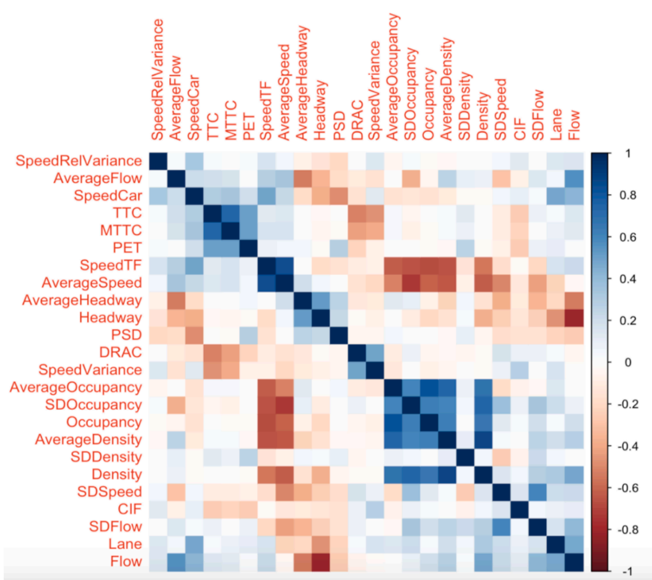

Fig. 11. Correlation plot of features.

volume increases, headway decreases. When occupancy increases, average speed decreases, followed by the increase of density.

For importance rankings, interactions and correlations can affect the stability of importance rankings (Darst et al., 2018). If SSMs are interdependent, small changes in the data or model might lead to significant changes in rankings. Correlation and interactions among variables may also lead to overfitting in high-dimensional SSMs data. Therefore, regularization techniques of Lasso regression is applied to handle multicollinearity and improve the stability of importance rankings. Cross validation and Bootstrapping tests across different scenarios have been conducted to evaluate how consistently certain SSMs are ranked as important. It shows that the top six most important variables are consistent in all scenarios. However, the latter variables vary from scenarios. It means that important variables truly captures SSM data characteristics and can indicate the occurrences of traffic conflicts.

This paper highlights the significance of recognising the varying efficacy of SSMs across different road environments, directly influencing their utility in real-world applications such as Advanced Driver Assistance Systems (ADAS) or roadside monitoring using smart infrastructure such as sensors developed by Vivacity Labs for promoting smarter, safe and sustainable cities. However, the study faced limitations due to the size of the dataset (1454 conflicts over more than 10 h of video recording, less than 0.1 % of total observations). Future investigations could strengthen the reliability of findings by expanding the scope to larger datasets. Additionally, further exploration could seek to identify the most effective SSMs in specific road environments. Furthermore, enriching the categorisation of road environments with more detailed classifications and considering diverse countries and levels of rurality could yield deeper insights into SSM performance across various contexts. Subsequent studies can incorporate additional enhancements, including assigning varying weights to individual ensemble models, establishing a hierarchical stacking framework through data partitioning and employing counterfactual causal inference for enhanced interpretability of variables.

Additionally, in this study, we tried as many SSMs as possible, but a few of them were not selected due to unable acquirement of the missing parameters by virtue of video analysis only. It possibly creates selection bias for choosing more longitudinal SSMs than cross sectional SSMs, which may be a potential limitation. For future research, sensors such as LiDar and Radar will be utilized to collect more sufficient traffic data and calculate corresponding SSMs.

The real-world width of the junction is measured and compared with the pixel width in the image to compute the scale parameter. This scale parameter is applied for all distance estimation as the drone height is fixed during data collection. The speed is calculated every second by using the difference in position per second. The smoothed distance is employed to ensure reliable speeds. The accuracy of the obtained speed values is computed by comparing them with speed measurements collected using a handheld speedometer simultaneously to drone data collection. The accuracy is presented in Table 2. Detection accuracy in different scenarios for UAV data.

Regarding the data balancing, random under-sampling of the non-conflict majority class was systematically employed to balance the data. Current real-time crash risk analyses primarily use undersampling methods to create equivalent ratios (commonly 1:1 or 1:4) between crash and non-crash samples through matched case-control techniques or by randomly discarding excess non-crash data (Abdel-Aty et al., 2004; Yang et al., 2018; Yu et al., 2020). Yu et al. (2020) conducted further investigation in finding proper crash to non-crash ratio in classification problem and 1:1 exhibits highest performance in all metrics: Sensitivity, FAR, AUC. Therefore, the number of conflicts in the dataset was calculated and the same amount of non-conflicts were selected to constitute 1:1 ratio of conflicts versus non-conflicts in the dataset. (See Table 7.)

Once the imbalanced dataset is scaled to a perfectly balanced dataset, differentiating between conflicts and non-conflicts data is easier since noises are filtered through during the under-sampling method.

The built stacked ensemble model has been tested in three scenarios: Vehicle based motorway, Infrastructure based motorway, and urban junctions. The detection accuracies cross validate each other. Due to distinct driving behaviours in conflict driving scenario and non-conflict driving scenario across different environments, conflict prediction accuracy may vary as well. For vehicle-basded data, it seems a bit overfit for traffic conflict prediction for having exceptionally high accuracy for both sensitivity and specificity. It probably ascribes to limited size of the dataset, which may also be the possible limitation of this study. In the future research, extra vehicle based dataset will be collected and cross validated with the proposed model to optimise robustness and test its suitability of in various scenarios.

Stacked ensemble model excels individual ensemble model in traffic conflict prediction which is the case invovles a mix of structured numerical data (SSMs) and unstructured data (temporal relationships and complex patterns). Component of the stacked ensemble model compensates each other to make complementary effects to capture temporal dependencies, to handle complex patterns, to achieve scalability to high-dimensional data, and to ensure model transferrability. Random Forest works well for tabular data with clear feature interactions and when model interpretability is important. SVM is useful for datasets where non-linear relationships and some temporal patterns can be captured through feature engineering. DNN is ideal for raw sequential or multi-modal data where temporal dependencies automatic feature learning are critical. GBM works well for classification and interpretability. Therefore, by combining these models, a more comprehensive approach to traffic conflict prediction can be achieved, leveraging the strengths of each method to address the specific characteristics of the data.

5. Conclusion

Different road environments possess different geometry and traffic characteristics, while vehicle behaviour and kinematics parameters exhibit significant variations in distinct road types. Safety surrogate measures are widely used safety indicators, albeit their significance and performance in different road environments are rarely investigated by researchers. This paper not only initiates a novel stacked ensemble learning model to predict traffic conflicts but also ranks SSM importance in different road scenarios during training, enabling more interpretable uses of black-box algorithms to avoid the trade-off between accuracy and interpretability. The novel stacked ensemble model achieves 99 %, 88 %, and 87.5 % prediction accuracy for vehicle-based motorway, urban junctions, and infrastructure-based motorway respectively. Ablation study results show that Gap time, CAI and DRAC are the three most significant SSMs for infrastructure-based motorway and urban junction scenarios. Conversely, significant variables for vehicle- differ considerably, with TTC, MTTC, PET, CIF, and PSD being the five most significant.

Future improvements for this study include splitting training data and constructing a weighted hierarchical stacked ensemble model to capture various aspects of traffic conflicts and assign different weights to individual models based on their respective performances. Additionally, recording more videos and identifying additional conflicts will expand the data pool, thereby enhancing the performance of the prediction models even further.

Table 7
Data Distribution.

Types	Conflict vs. Non-conflict Ratio	Ratio after undersampling
Vehicle based motorway	193:39017	193:193
Infrastructure based motorway	5181:147948	5181:5181
Urban junction	621:53070	621:621

CRediT authorship contribution statement

Bowen Cai: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation. **Léah Camarcat:** Writing – original draft, Validation, Software, Methodology. **Nicolette Formosa:** Investigation, Data curation, Conceptualization. **Mohammed Quddus:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

Ahmad, N., Wali, B., Khattak, A.J., 2023. Heterogeneous ensemble learning for enhanced crash forecasts – A frequentist and machine learning based stacking framework. J. Safety Res. 84, 418–434. https://doi.org/10.1016/j.jsr.2022.12.005.

Allen, B.L., Shin, B.T., Cooper, P.J., 1978. Analysis of traffic conflicts and collisions. Transp. Res. Rec.

An, R., Tong, Z., Ding, Y., Tan, B., Wu, Z., Xiong, Q., Liu, Y., 2022. Examining non-linear built environment effects on injurious traffic collisions: A gradient boosting decision tree analysis. J. Transp. Health 24, 101296. https://doi.org/10.1016/j.jth.2021.101296.

Arun, A., Haque, M.M., Bhaskar, A., Washington, S., Sayed, T., 2021. A systematic mapping review of surrogate safety assessment using traffic conflict techniques. Accid. Anal. Prev. 153, 106016. https://doi.org/10.1016/J.AAP.2021.106016.

Astarita V., Guido, Gi., Vitale, A., Giofré, V. (2012). A new microsimulation model for the evaluation of traffic safety performances. European Transport. 2-2.

Beest, D., Mes, S., Wilting, S., 2017. Improved high-dimensional prediction with Random Forests by the use of co-data. BMC Bioinf. 18, 584. https://doi.org/10.1186/s12859-017-1993-1.

Berk, R. (2016). Statistical Learning from a Regression Perspective. Third Edition. Springer. doi:10.1007/978-3-030-40189-4.

Bokaba, T., Doorsamy, W., Paul, B.S., 2022. A comparative study of ensemble models for predicting road traffic congestion. Appl. Sci. (switzerland) 12. https://doi.org/10.3390/app12031337.

Bonela, S.R., Kadali, B.R., 2022. Review of traffic safety evaluation at T-intersections using surrogate safety measures in developing countries context. IATSS Res. 46, 307–321. https://doi.org/10.1016/J.IATSSR.2022.03.001.

Darst, B., Malecki, K., Engelman, C., 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet. 19 (Suppl 1), 65. https://doi.org/10.1186/s12863-018-0633-8.

Diwakar, P., Landge, V., Jain, U., 2023. Evaluating the relationship between surrogate safety measures and traffic event severity in terms of human perception of danger: a perspective under indian traffic conditions. Appl. Sci. 13 (22), 12100. https://doi.org/10.3390/app132212100.

Federal Highway Administration (FHWA). (1989). U.S. Department of Transportation.

Formosa, N., Quddus, M., Ison, S., Abdel-Aty, M., Yuan, J., 2020. Predicting real-time traffic conflicts using deep learning. Accid. Anal. Prev. 136, 105429. https://doi.org/10.1016/j.aap.2019.105429.

Formosa, N., Quddus, M., Man, C., Timmis, A., 2023. Appraising machine and deep learning techniques for traffic conflict prediction with class imbalance. Data Sci. Transp. 5, 4.

Hastie, T., Friedman, J., Tibshirani, R., 2001. The elements of statistical learning. Springer Ser. Statist. https://doi.org/10.1007/978-0-387-21606-5.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. Boosting and Additive Trees. In: The Elements of Statistical Learning. Springer, New York, pp. 337–384.

He, Z., Qin, X., Liu, P., Sayed, M.A., 2018. Assessing surrogate safety measures using a safety pilot model deployment dataset. Transp. Res. Rec. 2672, 1–11. https://doi.org/10.1177/0361198118790861/ASSET/IMAGES/LARGE/10.1177_0361198118790861-FIG4.JPEG.

Hofmann, T., Schölkopf, B., Smola, A.J. (2008). Kernel methods in machine learning. 36 3 1171–1220. https://doi.org/10.1214/009053607000000677.

Hofmann, T., Schölkopf, B., Smola, A.J., (2008). Kernel methods in machine learning. https://doi.org/10.1214/009053607000000677 36, 1171–1220. https://doi.org/10.1214/009053607000000677.

Hydén, C., 1987. The Development of a Method for Traffic Safety Evaluation: The Swedish Traffic Conflicts Technique; Bulletin Lund Institute of Technology. Department of Traffic Planning and Engineering, Lund, Sweden.

Johnsson, C., Laureshyn, A., Dágostino, C., 2021. A relative approach to the validation of surrogate measures of safety. Accid. Anal. Prev. 161, 106350. https://doi.org/10.1016/j.aap.2021.106350.

Kalule, R., Abderrahmane, H.A., Alameri, W., Sassi, M., 2023. Stacked ensemble machine learning for porosity and absolute permeability prediction of carbonate rock plugs. Sci. Rep. 13. https://doi.org/10.1038/s41598-023-36096-2.

Li, Z., Li, Y., Liu, P., Wang, W., Xu, C., 2014. Development of a variable speed limit strategy to reduce secondary collision risks during inclement weathers. Accid. Anal. Prev. 72, 134–145.

Lu, C., He, X., van Lint, H., Tu, H., Happee, R., Wang, M., 2021. Performance evaluation of surrogate measures of safety with naturalistic driving data. Accid. Anal. Prev. 162, 106403. https://doi.org/10.1016/J.AAP.2021.106403.

Mahmud, L., Ferreira, M.S., Hoque, A., Tavassoli. Application of proximal surrogate indicators for safety evaluation: a review of recent developments and research needs.

Marcillo, P., Valdivieso Caraguay, Á.L., Hernández-álvarez, M., 2022. A Systematic literature review of learning-based traffic accident prediction models based on heterogeneous sources. Appl. Sci. 2022 12, 4529. https://doi.org/10.3390/APP12094529.

Mason, L., Bartlett, P., Baxter, J., Frean, M., 1999. Boosting Algorithms as Gradient Descent. Advances in Neural Information Processing Systems 12.

Minderhoud, M.M., Bovy, P.H.L., 2001. Extended time-to-collision measures for road traffic safety assessment. Accid. Anal. Prev. 33, 89–97. https://doi.org/10.1016/S0001-4575(00)00019-1.

Nadimi, N., Amiri, A.M., Sadri, A., 2021. Introducing novel statistical-based method of screening and combining currently well-known surrogate safety measures. Transp. Lett. 14 (4), 385–395. https://doi.org/10.1080/19427867.2021.1874184.

Natekin, A., Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics.

Niyogisubizo, J., Liao, L., Zou, F., Han, G. Nziyumva, E., Li, B., Lin, Y. (2023). Predicting Traffic Crash Severity Using Hybrid of Balanced Bagging Classification and Light Gradient Boosting Machine. 27, 1 79–101.

Okamura, M., Fukuda, A., Morita, H., Suzuki, H., Nakazawa, M., 2011. Impact Evaluation of A driving Support System on Traffic Flow by Microscopic Traffic Simulation. In: 3rd International Conference on Road Safety and Simulation.

Orsini, F., Gecchele, F., Rossi, R., Gastaldi, M., 2021. A conflict-based approach for real-time road safety analysis: Comparative evaluation with crash-based models. Accid. Anal. Prev. 161, 106382. https://doi.org/10.1016/j.aap.2021.106382.

Ozbay, K., Yang, H., Bartin, B., Mudigonda, S., 2008. Derivation and validation of new simulation-based surrogate safety measure. Transp. Res. Rec. 105–113. https://doi.org/10.3141/2083-12.

Saccomanno, F.F., Cunto, F., Guido, G., Vitale, A., (2008). Comparing Safety at Signalized Intersections and Roundabouts Using Simulated Rear-End Conflicts. https://doi.org/10.3141/2078-12 90–95. https://doi.org/10.3141/2078-12.

Tak, S., Kim, S., Lee, D., Yeo, H., 2020. A comparison analysis of surrogate safety measures with car-following perspectives for advanced driver assistance system. J. Adv. Transp. 2018. https://doi.org/10.1155/2018/8040815.

Tang, J., Alelyani, S., Liu, H., 2014. Feature Selection for Classification: A Review. Data Classification: Algorithms and Applications.

Tang, J., Liang, J., Han, C., Li, Z., Huang, H., 2019. Crash injury severity analysis using a two-layer Stacking framework. Accid. Anal. Prev. 122, 226–238. https://doi.org/10.1016/j.aap.2018.10.016.

Tajiri, Y., Yabuwaki, R., Kitamura, T., Abe, S., 2010. Feature Extraction Using Support Vector Machines. In: Wong, K.W., Mendis, B.S.U., Bouzerdoum, A. (Eds.), Neural Information Processing. Springer, Berlin, Heidelberg, p. 6444. https://doi.org/10.1007/978-3-642-17534-3_14.

Thissen, U., Brakel, R., Weijer, A., Melssen, W., Buydens, L., 2003. Using support vector machines for time series prediction. Chemom. Intel. Lab. Syst. 69 (1–2), 35–49. https://doi.org/10.1016/S0169-7439(03)00111-4.

Thurnhofer-Hemsi, K., López-Rubio, E., Molina-Cabello, M.A., Najarian, K., (2020). Radial basis function kernel optimization for Support Vector Machine classifiers.

Thurnhofer-Hemsi, K., López-Rubio, E., Molina-Cabello, M.A., Najarian, K. (2020). Radial basis function kernel optimization for Support Vector Machine classifiers.

Uno, N., Iida, Y., Itsubo, S., Yasuhara, S., 2002. A microscopic analysis of traffic conflict caused by lane-changing vehicle at weaving section. Accid. Anal. Prev. 33, 89–97.

Van der Horst, A., 1984. The ICTCT Calibration Study at Malmoe: A Quantitative Analysis of Video Recording. National Academy of Sciences, Washington DC USA.

Vogel, K., 2003. A comparison of headway and time to collision as safety indicators. Accid. Anal. Prev. 35, 427–433. https://doi.org/10.1016/S0001-4575(02)00022-2.

Wang L., Wu J., Li R., Song Y., Zhou J., Rui X., Xu H. (2021). A weight assignment algorithm for incomplete traffic information road based on fuzzy random forest method, Symmetry. 13(9). doi: 10.3390/sym13091588.

Wang, C., Stamatiadis, N., 2014. Evaluation of a simulation-based surrogate safety metric. Accid. Anal. Prev. 71, 82–92. https://doi.org/10.1016/J.AAP.2014.05.004.

Wang, C., Xie, Y., Huang, H., Liu, P., 2021b. A review of surrogate safety measures and their applications in connected and automated vehicles safety modeling. Accid. Anal. Prev. 157, 106157. https://doi.org/10.1016/j.aap.2021.106157.

World Health Organization, (2023). Global status report on road safety.

Xiao, J., 2019. SVM and KNN ensemble learning for traffic incident detection. Phys. A: Statist. Mechan. Appl. 517, 29–35. https://doi.org/10.1016/j.physa.2018.10.060.

Yan M., Shen Y. (2022). Traffic accident severity prediction based on random forest, Sustainability. 14(3). doi: 10.3390/su14031729.

Yu, R., Wang, Y., Zou, Z., Wang, L., 2020. Convolutional neural networks with refined loss functions for the real-time crash risk analysis. Transp. Res. Part C Emerg. Technol. 119, 102740. https://doi.org/10.1016/J.TRC.2020.102740.

Zhang, H., Gao, S., Wang, P., 2022. Stacking-based ensemble learning method for the recognition of the pedestrian crossing intention. J. Adv. Transp. 2022. https://doi.org/10.1155/2022/6601014.

Zhang, Y., Haghani, A., 2015. A gradient boosting method to improve travel time prediction. Transp. Res. Part C: Emerg. Technol. 58 (Part B), 308–324. https://doi.org/10.1016/j.trc.2015.02.019.

Zhao, H., Li, X., Cheng, H., Zhang, J., Wang, Q., Zhu, H., 2022. Deep learning-based prediction of traffic accidents risk for Internet of vehicles. China Commun. 19 (2), 214–224. https://doi.org/10.23919/jcc.2022.02.017.

Zheng, L., Sayed, T., Mannering, F., 2021. Modeling traffic conflicts for use in road safety analysis: A review of analytic methods and future directions. Anal. Methods Accid. Res. 29, 100142. https://doi.org/10.1016/J.AMAR.2020.100142.

Zhong, W., Du, L., 2023. Predicting traffic casualties using support vector machines with heuristic algorithms: a study based on collision data of urban roads. Sustainability 15 (4), 2944. https://doi.org/10.3390/su15042944.