Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review

Yasir Ali[a,*], Fizza Hussain[b], Md Mazharul Haque[b]

[a] *School of Architecture, Building, and Civil Engineering, Loughborough University, Leicestershire LE11 3TU, United Kingdom*
[b] *Queensland University of Technology, School of Civil & Environment Engineering, Faculty of Engineering, Brisbane 4001, Australia*

ARTICLE INFO

Keywords:
Safety
Crash frequency
Real-time
Injury severity
Machine learning
Systematic review

ABSTRACT

Accurately modelling crashes, and predicting crash occurrence and associated severities are a prerequisite for devising countermeasures and developing effective road safety management strategies. To this end, crash prediction modelling using machine learning has evolved over two decades. With the advent of big data that provides unprecedented opportunities to better understand the crash mechanism and its determinants, such efforts will likely be accelerated. To gear these efforts, understanding state-of-the-art machine learning-based crash prediction models becomes paramount to summarise the lessons learned from past efforts, which can assist in developing robust and accurate models. This review paper aims to address this gap by systematically reviewing the machine learning studies on crash modelling. Models are reviewed from three aspects of the application: (a) crash occurrence (or real-time crash) prediction, (b) crash frequency prediction, and (c) injury severity prediction. Further, model intricacies that impact model performance are identified and thoroughly reviewed. This comprehensive review highlights specific gaps and future research needs in three aforementioned model applications, such as improper selection of non-crash events for crash occurrence models, the inability of future forecasting of crash frequency models, and inconsistency in injury severity classes. Critical research needs relating to model development, evaluation, and application are also discussed. This review envisages methodological advancements in machine learning models for crash prediction modelling and leveraging big data to better link crashes with its determinants.

1. Introduction

1.1. Problem definition

According to World Health Organisation, every year, about 1.3 million people die in road traffic crashes, with 20 to 50 million suffering non-fatal injuries (WHO, 2023). These crashes incur disability to many people involved in crashes, costing most countries 3% of their gross domestic product. Further, it has been reported that 54% of vulnerable road users (i.e., pedestrians, cyclists, and motorcyclists) are involved in fatal crashes. These crashes mainly involve young people (aged between 5 and 29 years) and males (WHO, 2023). Similarly, motorcyclists have significantly higher chances of being involved in a fatal crash than car occupants (Lin and Kraus, 2008). These alarming statistics call for improving road safety, which is also aligned with the goal of the World Health Organisation. To this end, a deeper and sound understanding of the crash mechanism and its link to underlying determinants is critical and motivates this study.

A common way to understand the crash mechanism is to develop models that can concurrently link crashes with its contributing factors, including but not limited to roadway geometric features, traffic characteristics, vehicle design, driver characteristics and weather. Recognising the importance of crash prediction modelling, many studies have developed models to better understand road safety. With a growing interest from both researchers and practitioners in advancing the field of road safety, it is paramount to critically evaluate the state-of-the-art crash modelling and appraise future research directions, which form the basis of this study.

1.2. Research gap

Crash prediction modelling is generally performed using statistical (and econometric) and machine learning models, with the former class having a long history and being reviewed comprehensively (see

[*] Corresponding author.
E-mail addresses: y.y.ali@lboro.ac.uk (Y. Ali), fizza.hussain@hdr.qut.edu.au (F. Hussain), m1.haque@qut.edu.au (M.M. Haque).

https://doi.org/10.1016/j.aap.2023.107378
Received 23 June 2023; Received in revised form 29 October 2023; Accepted 8 November 2023
Available online 15 November 2023

Table 1
A summary of review studies on machine learning for crash modelling.

Study	Crash models			Years covered	Studies considered	Model intricacies covered	Issues discussed
	Occurrence	Frequency	Injury severity				
Silva et al. (2020)	×	✓	✓	2003–2020	26*	×	Methodological issues and selection of explanatory variable
Angarita-Zapata et al. (2021)	×	×	✓	2011–2021	53	×	Bibliometric review and comparing machine learning and auto machine learning for three Colombian cities
Wen et al. (2021)	×	×	✓	2000–2020	88*	×	Data imbalance and under-reporting, sensitivity analysis, model goodness of fit, spatiotemporal correlations, casualty, and transferability
Santos et al. (2022)	×	×	✓	2001–2021	56	×	Bibliometric review, learning parameters, predictions, and explanatory variables.
Current study	✓	✓	✓	1997–2023	213	✓	See Section 5

* These numbers are obtained by counting the number of studies from the tables/figures provided in the respective studies.

Mannering and Bhat (2014) and Mannering et al. (2016) for more details). Comparatively, the application of machine learning for crash modelling has gained traction over the past two decades, mainly due to its superior performance for predictions. Among several characteristics of machine learning models (a detailed discussion to follow in the next section), these models are flexible, require no prior assumption about data distribution, and can handle missing values (Tang et al., 2019). With significant breakthroughs in artificial intelligence leading to several (and much needed) advancements in machine learning, this review study focusses on machine learning models specifically applied for crash prediction modelling.

Despite the increasing popularity of machine learning for crash prediction modelling, our understanding of properly developing and rigorously testing machine learning models remains elusive mainly because of the lack of a proper framework for model development and benchmarking its performance. To this end, a comprehensive review of machine learning models, which scrutinises and summarises notable efforts and achievements in crash modelling aspects (i.e., crash occurrence, crash frequency, and injury severity), pinpoints challenges and issues in existing research, and inspires more sophisticated and rigorously tested models, is long overdue. Along this line, a few studies have attempted to review existing machine learning models,[1] especially for injury severity (Santos et al., 2022, Angarita-Zapata et al., 2021, Wen et al., 2021, Silva et al., 2020). The review offered in our study, however, is more comprehensive and detailed. To support this argument, Table 1 summarises the efforts of past review papers, and evidently, all these papers majorly focus on injury severity models, except Silva et al. (2020), who briefly discussed crash frequency models as well. None of the earlier reviews considered models specifically developed for crash occurrence prediction (or real-time crash prediction models), which is a significant research gap.

The reviews of Santos et al. (2022) and Angarita-Zapata et al. (2021) are bibliometric analyses, providing a picture of how machine learning models for injury severity have evolved over the years and which algorithms have been used, whilst less attention has been paid to identify crash modelling challenges and future research needs. These shortcomings were partly addressed by Wen et al. (2021) and Silva et al. (2020), who discussed some modelling issues, including explanatory variables, data imbalance and under-reporting, sensitivity analysis, model goodness of fit, spatiotemporal correlations, casualty, and transferability. However, several other (important) intricacies of machine learning models — as identified and discussed later in this paper — remain largely undiscussed. For instance, hyperparameter tuning and its effects on model performance, different training and test proportions, counterintuitive model comparison output (e.g., a logistic regression model outperforming a machine learning model), and overfitting are some of the many issues discussed in the current review paper. Further, none of the existing reviews discuss the models and their applicability for assessing safety of emerging vehicle technologies, such as connected and automated vehicles. Finally, although past review studies identified some key issues like unobserved heterogeneity, they did not provide guidelines on how to address such issues.

1.3. Objective and contributions

To overcome the research gaps mentioned above, this study comprehensively reviews machine learning-based crash prediction models specifically developed for crash occurrence (or real-time) prediction, crash frequency prediction, and injury severity prediction.

The contribution of this paper is twofold. First, this study comprehensively reviews machine learning models from two perspectives: (a) models specifically developed for the three aspects of crash prediction modelling, i.e., crash occurrence, frequency, and injury severity, and (b) general model intricacies. To the best of the author's knowledge, this study is the first attempt to comprehensively review machine learning models at a granular level, providing detailed insights into model intricacies, which are largely ignored in the literature. Second, this study highlights specific research needs for different categories of models and general research needs, which apply to all machine learning models. These research needs also entail emerging vehicle technologies, such as connected and automated vehicles, as their probe data provide opportunities to better understand the dynamics of road safety.

1.4. Outline

The rest of the paper is organised as follows. Section 2 describes the study framework and systematic review process, with a brief bibliometric analysis and the study's scope. Section 3 provides an overview of machine learning models for three aspects of crash prediction modelling: crash occurrence, crash frequency, and injury severity. Whilst Section 4 reviews model intricacies from different perspectives, Section 5 presents general and specific future research needs categorised into model development, evaluation, and application aspects. Finally, Section 6 concludes this review study.

2. Study framework

The overall study framework is presented in Fig. 1. The review is conducted in four stages, with Stages 1 to 3 summarising existing studies and Stage 4 eliciting future research directions. Stage 1 classifies the models based on the crash aspect, i.e., crash occurrence, crash frequency, and injury severity predictions. Stage 2 explains different

[1] The detailed definitions of crash occurrence, frequency, and injury severity models are defined in Section 3. Briefly, crash occurrence models predict whether a crash will occur or not, crash frequency models predict aggregated crashes, and crash injury severity models predict different injury classes.

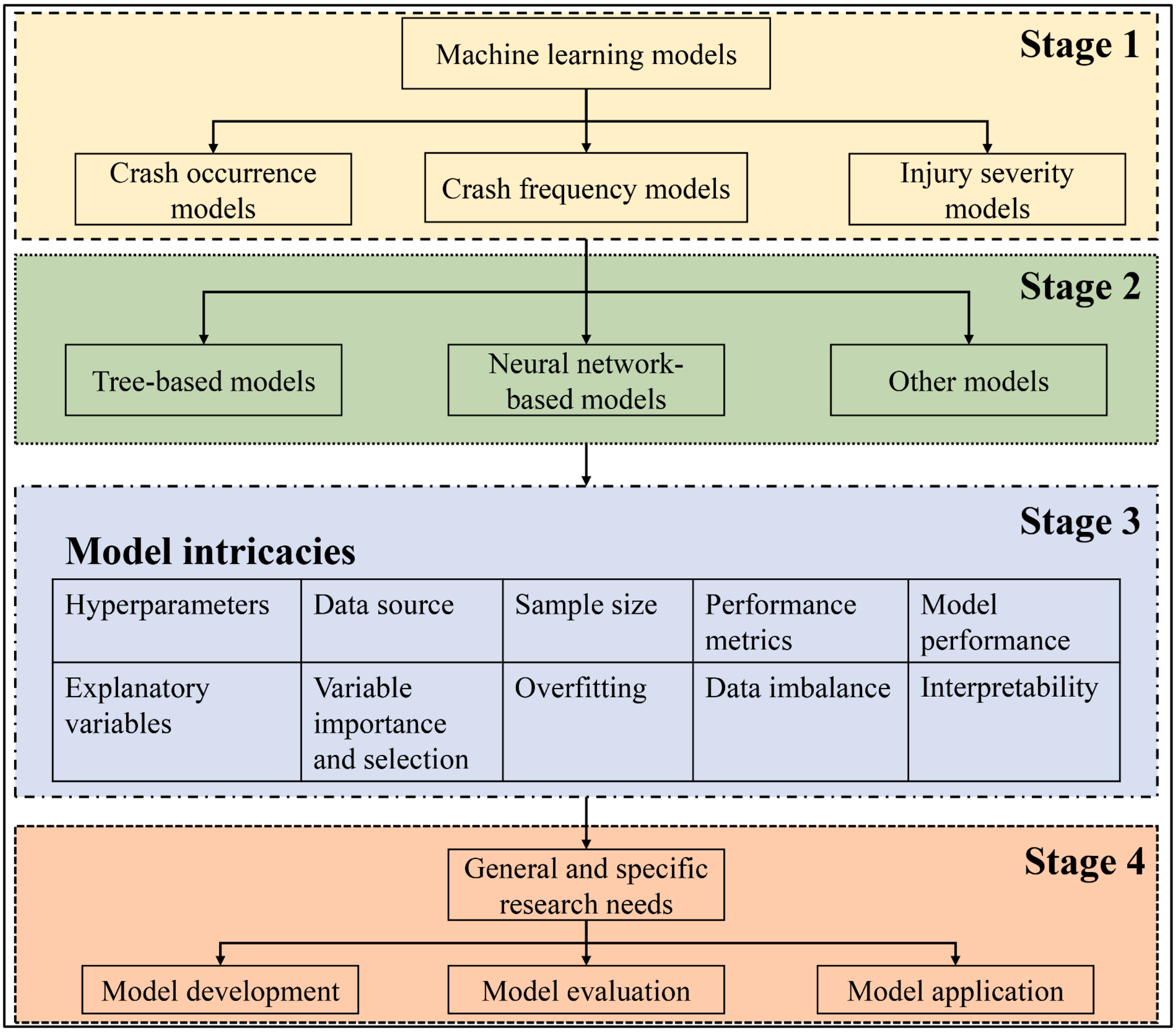

Fig. 1. The overall review framework.

machine learning models based on their underlying methodology, such as tree-based (e.g., decision tree, random forest, etc.) and neural network-based (e.g., artificial neural network, convolutional neural network, and recurrent neural network). Stage 3 reviews modelling intricacies, including but not limited to hyperparameters selection, data source, sample size, interpretability, transferability, and so on. Finally, Stage 4 provides specific and general future research directions. Each of these stages is covered and discussed in the following sections.

Note that Stages 1–3 (mapped to Sections 3 and 4) provide a review of existing studies, whereas Section 5 provides future research directions based on the thorough review presented in Sections 3 and 4. This classification makes it easier to locate the relevant information.

2.1. *Study scope*

As mentioned above, the scope of this study is limited to three machine learning-based crash prediction modelling aspects: crash occurrence (or real-time), crash frequency, and injury severity. To systematically search and review the existing literature, this study adopted the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Sarkis-Onofre et al., 2021, Liberati et al., 2009). Since this study deals with machine learning techniques, it is important to clarify the term “prediction” in the context of this study. Prediction in the context of machine learning-based crash aspect modelling refers to accurately predicting different crash aspects using unseen (or testing) datasets during model development. It is worth noting that predicting should not be confused with forecasting, whereby the aim is to predict future injury severities/crash frequency/crash occurrences conditioned on a lot of unknown factors (e.g., age, gender, vehicle type, impact speed, collision manner, weather, etc.). Finally, whilst predicting crash aspects is helpful to devise countermeasures for reducing crash occurrence and resulting injury severities, these predictions are only valid for certain time periods and within a geographic area or a road facility, and their transferability needs to be thoroughly assessed (see more information in Section 5.2.7).

Further, this study defined relevant and precise eligibility criteria using the Population, Intervention, Comparator group, Outcome, Study (PICOS) design approach (Methley et al., 2014). The ‘Population’ of studies included all peer-reviewed studies in journals and conferences in the last 26 years (1997–2023) concerning the modelling of different crash aspects. The ‘Intervention’ ensured that studies on predicting three crash modelling aspects were selected. The ‘Comparator group’ included studies comparing machine learning models with statistical (and econometric) models. The ‘Outcome’ of machine learning studies varied by the intervention (i.e., occurrence, frequency, and injury severity) and was recorded accordingly. Finally, the ‘Study design’ considered machine learning studies from both categories, i.e., original research and case studies demonstrating the application of machine learning in highway safety.

A thorough and systematic literature search was performed on

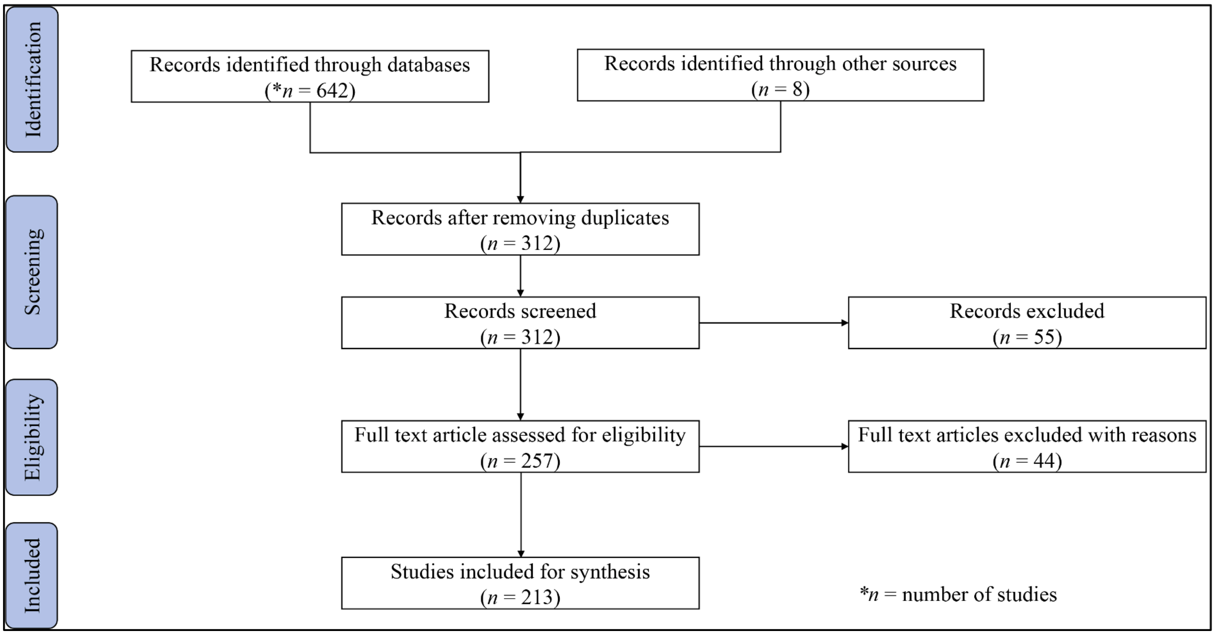
Identification
Screening
Eligibility
Included
Records identified through databases (*n = 642)
Records identified through other sources (n = 8)
Records after removing duplicates (n = 312)
Records screened (n = 312)
Records excluded (n = 55)
Full text article assessed for eligibility (n = 257)
Full text articles excluded with reasons (n = 44)
Studies included for synthesis (n = 213)
*n = number of studies

Fig. 2. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram for the systematic review.

several known scientific databases, such as Scopus, Web of Science, Compendex and Inspec (via Engineering Village), Science Direct, Pubmed, Transport Research International Documentation (TRID), IEEE Xplore, Taylor & Francis, Springer, Google Scholar, Mendeley, and University library databases. For the sake of comprehensiveness, several keywords and their combinations were used in the databases mentioned above. These keywords were divided into two classes: (1) machine learning, including machine learning, deep learning, decision tree, classification and regression trees, support vector machine, neural networks, random forests, Bayesian networks, and so on; and (2) crash modelling, including crash occurrence, real-time crash prediction, crash severity, crash injury, and injury severity, crash frequency and so on. Note that Boolean operators like 'AND'/ 'OR' were used for searching the relevant studies. The selection of keywords was driven by the terms commonly used and found in the literature. The snowballing method (Wohlin, 2014) was used from the initially identified articles to find the missing relevant papers.

A comprehensive search was performed covering the studies from the last 26 years (1997–2023) to highlight the evolution of machine learning since its first application for crash prediction modelling. As a result of this search strategy, a total of 650 relevant studies were obtained. Note that the search for 2023 was limited till February 12th, 2023. Fig. 2 displays the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram of the number of studies obtained and included at each stage. In the initial screening, duplicate studies and papers unrelated to crash prediction modelling were omitted, resulting in 312 studies. These studies were further mapped with the inclusion criteria of PICOS, and some selected inclusion criteria for considering a study are:

- studies only in the English language;
- studies published in peer-reviewed journals and conferences;
- studies employing any machine learning model either as the main focus of the study or for comparison;

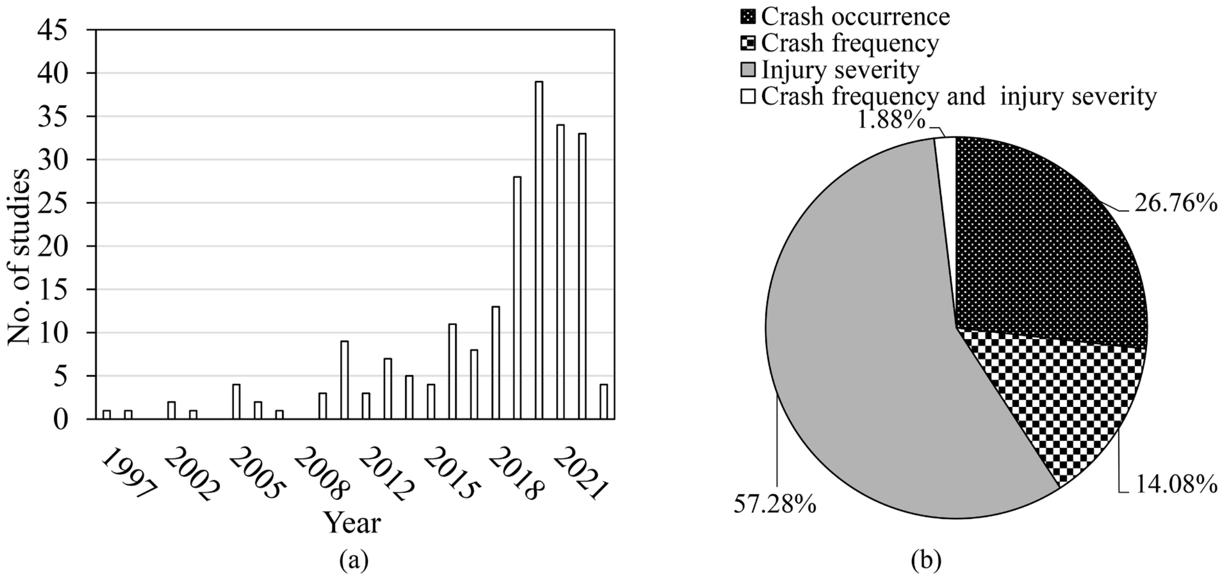

Fig. 3. Schematics for bibliometric analysis: (a) yearly publications trend and (b) proportion of publications for different crash modelling aspects.

Table 2
A summary of the working mechanism of different crash model types.

Characteristic	Crash modelling aspect		
	Occurrence	Frequency	Injury severity
Problem	Binary classification (yes/no as the dependent variable)	Regression (number of crashes as the dependent variable)	Binary or multiclass classification (yes/no in binary, injury severity levels in multiclass, e.g., property damage only, minor injury, severe injury, and fatal injury)
Performance measure(s)	Confusion matrix (and its elements, i.e., true positive, false positive, true negative, false negative), precision, recall, accuracy, specificity, F_1 score, G-mean, and Kappa statistics	Mean absolute error, root mean squared error, mean absolute deviation, mean squared prediction error, mean absolute prediction error, R^2, and validation plot	Confusion matrix (and its elements, i.e., true positive, false positive, true negative, false negative), precision, recall, accuracy, specificity, F_1 score, G-mean, and Kapp statistics
Data	Police reported crash data and non-crash events generated using traffic and other information	Police reported aggregated crash data	Police reported crash data, containing injury severity scores
Class imbalance	Exists as crashes are rare and non-crash events are frequent	Not applicable	Exists as some levels of injury classes are often underreported whilst others are overrepresented

- studies aimed at modelling crash occurrence (or real-time crash), crash frequency, and injury severity.

Similarly, studies were excluded if machine learning was applied with a different motivation than the crash prediction, such as obtaining high-order interactions (Ali et al., 2021) and explaining unobserved heterogeneity for modelling driving behaviour (Ali et al., 2022). Theses and dissertations were excluded if papers from such documents were published and considered. With these inclusion and exclusion criteria, a total of 213 studies were identified and reviewed in this study.

Note that the scope of the study covers all types of roadways (e.g., rural and urban arterials, motorways/freeways/expressways, mountainous roads, signalised and unsignalised intersections, and others). Similarly, all types of road users (e.g., vehicles, pedestrians, motorcycles, bicycles, and so on) are considered in this study. To summarise, any study using machine learning for predicting any crash aspect is considered in this review.

2.2. *Bibliometric analysis*

Fig. 3 displays the bibliometric analysis of the considered studies. Fig. 3 (a) illustrates that the first study of machine learning application for crash prediction modelling was published in 1997, with a gradual increase in the number of such studies till 2017. Fig. 3 (a) shows an upward trend thereafter, reflecting the increased interest in applying machine learning for highway crash prediction modelling in recent years.

Fig. 3 (b) indicates the number of studies for different crash aspects, and it can be observed that the number of machine learning-based injury severity models is the highest. Similarly, machine learning-based crash occurrence (or real-time crash) prediction models have also received significant attention but have never been reviewed before, deserving consideration in this study. Finally, the combined modelling of crash injury severity with crash frequency using real-time traffic data has comparatively received less attention, which is clearly disproportionate to its importance (more discussion in Section 5).

This bibliometric analysis indicates the temporal scope that extends from 1997 to 2023, covering historical perspectives as well as the rapid and transformative developments in machine learning over the past decade. Over time, algorithms and architectures have drastically evolved because of computational power and data availability, and this substantial landscape change has been fully covered in this study, whereby the shift from conventional methods to advanced methods is demonstrated through some representative studies.

3. A review of machine learning models

This section reviews machine learning models applied for crash modelling and divides them into two groups: (a) machine learning models based on the type of crash aspect modelled and (b) machine learning models based on modelling methodology.

3.1. *Crash aspect-based classification*

Various models are developed in the literature that can be broadly classified into three groups based on the type of crash modelling aspect: (a) crash occurrence (or real-time crash) prediction models, (b) crash frequency prediction models, and (c) injury severity prediction models. Whilst a detailed discussion on model intricacies is presented in the next section, general model characteristics of each type of crash aspect are presented in Table 2.

3.1.1. *Crash occurrence prediction models*

Crash occurrence (or real-time crash) prediction models are typically binary classification (crash vs. no crash) models aimed at predicting the likelihood of crash occurrence given a set of traffic conditions. Whilst the crashes are obtained from historical crash records, non-crash events are usually randomly selected and linked with real-time driving and environmental conditions. How to select appropriate non-crash events and how many non-crash events corresponding to one crash should be selected are some of the many questions discussed in the next section.

Several machine learning models are developed for real-time crash occurrence prediction using decision trees and neural networks. For instance, Li and Abdel-Aty (2022b) developed a real-time crash occurrence prediction model using temporal attention–based deep learning and trajectory fusion. Trajectory data were collected using the automatic vehicle locator and Lytx DriveCam. The modelling data were prepared as follows. Traffic conditions (5–10 mins) before the crash were considered contributory factors, and the time slices were used to label a crash or non-crash. For instance, if a crash occurred at 11:50, the crash occurrence status from 11:40 to 11:45 was marked as 1 (crash occurred), indicating that a crash would happen in the next 5–10 mins. Otherwise, the crash occurrence status was marked as 0 (non-crash). The model showed a reasonable performance in crash occurrence prediction. Karim et al. (2022) developed a dynamic spatial–temporal attention network model for early anticipation of a traffic crash utilising dashcam video data. A Gated Recurrent Unit was trained jointly with the attention modules to predict the probability of a future accident, which was confirmed by using two benchmark datasets. Several other studies applied advanced methodologies for real-time crash occurrence prediction, e.g., Fang et al. (2022), Lin et al. (2021), He et al. (2021), Li et al. (2020). However, our understanding remains elusive on (a) how long the time window should be considered for a non-crash event?, (b) whether the time window has any effect on model performance?, and (c) how this method of data preparation performs compared to other methods like the case-control approach?. These are some of the many research questions discussed in Section 5.

3.1.2. *Crash frequency prediction models*

Crash frequency prediction models aim to predict the frequency of crashes that occur in a given time. As such, crash frequency models are regression models, unlike classification-based crash occurrence models. In these models, the dependent variable is the aggregated number of crashes, which is modelled as a function of several explanatory variables, such as annual average daily traffic and other road traffic characteristics. The output of these models (i.e., crash frequency) is compared with the observed crashes using performance measures, such as mean absolute error and root mean square error. Li et al. (2008) used support vector machines to predict crash frequency on rural frontage roads in Texas with different sample sizes and performed a comparative analysis with a conventional negative binomial model. Explanatory variables considered in this study were the length of road segments, average daily traffic, right-shoulder length, and lane width. The support vector machine model showed better results in predicting the number of crashes than a negative binomial model. Interestingly, with an increase in model training sample size, the model performance deteriorated, which is counterintuitive and remains unanswered in the study (Li et al., 2008). A recent study on crash frequency applied conditional generative adversarial networks for identifying crash hotspots (Zarei et al., 2023). Using a series of simulation, results confirmed the efficacy of the proposed method for predicting crash frequency.

Although crash frequency prediction modelling is relatively straightforward, data (quality and quantity) and its impact on model performance is a long puzzling problem (more discussion on this issue in crash frequency models in Section 5).

3.1.3. *Injury severity prediction models*

Injury severity prediction is also a classification task, with different studies framing their problem as binary classification (fatal injury or not) or multiclass classification (3, 4, or 5 injury severity levels). The dataset used for this type of crash modelling is primarily collected by police, reporting various information, such as driver characteristics (e.g., gender and age), time of the day, weather conditions, and many others. Assi (2020) proposed a hybrid system using principal component analysis with multilayer perceptron neural networks and support vector machines to predict injury severity. The use of principle component analysis assisted in explanatory variable selection, and the developed model was found to predict two classes of injury severity (fatal/serious and slight injury) with reasonable accuracy. This study demonstrated that feature selection improves model performance; however, the influence of various feature selection techniques is yet to be investigated. Another study presented a transparent deep machine learning framework for predicting crash injury severity on different road locations (Sattar et al., 2023), and compared three different approaches, namely plain vanilla multi-layer perceptron (MLP) using Keras, MLP with embedding layers, and TabNet. Their study reported that TabNet may be better suited for complex data structures, whereas for real-time applications where computational efficiency is required, MLP using Keras should be preferred.

3.2. *Modelling methodology-based classification*

This study broadly classifies machine learning-based crash prediction models into three groups based on modelling methodology: (a) decision tree-based models, (b) neural network-based models, and (c) other models (e.g., support vector machine, Bayesian network). All these model classes are summarised in the ensuing subsections.

3.2.1. *Decision tree-based models*

Decision tree-based models graphically illustrate internal structures (or nodes), which can be easily understood to interpret model results (Chang and Wang, 2006). Decision trees are often subject to overfitting issues, which are handled using pruning that prevents overfitting by cutting the branches of the tree that do not significantly contribute to prediction performance (Chang and Wang, 2006). Contrary to these advantages, decision trees are biased to select categorical variables with higher categories for internal nodes, resulting in poor predictive performance and generalisation capabilities. Moreover, decision tree structures are often unstable and susceptible to drastic change if different strategies are applied to obtain training and testing datasets (Jeong et al., 2018). Finally, a common notion is that decision trees possess poor predictive performance compared to state-of-the-art methods like deep learning (Wen et al., 2021). However, Candefjord et al. (2021) found that, in some instances, decision trees can outperform deep learning models. Commonly used decision tree models are explained below with one representative study of their application.

Classification and Regression Trees (CART) consider categorical and continuous variables as input or output and apply a recursive binary splitting strategy. These models are developed in three steps: tree growing, pruning, and determining the right size of the pruned trees (Chang and Chen, 2005). In the tree-growing step, recursive partitioning of the target (or response) variable is performed to minimise impurity in the terminal node, whereas, in the second step, pruning is performed to create simpler trees by considering important nodes. Finally, a right-sized tree is determined by misclassification cost on a new (or independent) dataset to avoid overfitting. To this end, data are often divided into two sets: training (or learning) and testing. The tree size is usually determined when misclassification costs reach a minimum for both learning and testing datasets (Chang and Chen, 2005). CART has been applied for modelling the three aspects of the crash mentioned above and has shown reasonable predictive performance. For instance, Chang and Chen (2005) modelled freeway crash frequency with CART and compared the results with a negative binomial regression model, with CART providing marginally higher prediction accuracy (52.6 % compared to 52.3 % for a negative binomial model).

Other than CART, several variants of decision tree models are also applied, including J48, ID3, C4.5, and C5. Unlike CART, which combines specific categories (three or more) of independent variables and cannot distinguish the impact of individual categories on the target variable, J48, ID3, C4.5, and C5 are free from this binary split restriction. Contrary to CART, which employs the Gini index as the splitting criterion, J48, ID3, and C4.5/C5 use normalised information gain, information gain, and information gain ratio, respectively. De Oña et al. (2013) compared the performance of CART, ID3, and C4.5 for injury severity prediction, and the results revealed that CART has relatively better predictive capabilities (receiver operating characteristic area of 0.57).

Ensemble learning methods: To address the performance limitations of simple decision trees like CART, ensemble learning methods are applied for crash prediction modelling, aiming to combine multiple weak classifiers to improve predictive capabilities. Ensemble learning methods can be broadly classified into bagging and boosting, and the models based on these methods are discussed below.

Random Forest is a widely used ensemble learning (and bagging) method (Breiman, 2001), combining several decision trees with different characteristics and predictive capabilities to obtain relatively high predictive power. A random forest model trains a set of decision tree classifiers, each classifier receiving samples through bagging. During the training, the nodes of each classifier are split based on a randomly selected subset of explanatory variables. With trained classifiers and input variables, each classifier votes for the target variable and the final classification is obtained based on the majority of votes. Compared to simple decision trees, random forest models (a) possess higher predictive performance, (b) are insensitive to noise, outliers, and overfitting, and (c) can provide the importance of explanatory variables (or features) in the model. Random forest is frequently used for crash prediction modelling, with applications in crash occurrence, crash frequency, and injury severity predictions. Zhang et al., (2022b) developed a random forest model for real-time crash occurrence prediction on freeways using crowdsourced probe vehicle data. This model outperformed several competing models like logistic regression, extreme gradient boosting, and support vector machine.

In another class of ensemble learning, *Adaptive Boosting* (*AdaBoost*), *Gradient Boosted Decision Tree*, and Extreme Gradient Boosting models are developed following boosting techniques. AdaBoost models are adaptive as they adjust subsequent weak classifiers that are more prone to misclassification. In other words, AdaBoost models decrease and increase the weight of correctly and incorrectly classified samples, respectively. As such, the final predictions are obtained by a weighted majority vote of individual classifiers' predictions. It has been applied for injury severity and real-time crash occurrence predictions. For instance, Almamlook et al. (2019) employed an AdaBoost model for injury severity prediction and compared its performance with several competing models, including Naïve Bayes and Random Forests. Results showed that the AdaBoost model outperformed Naïve Bayes but was inferior to Random Forests.

Contrasting to AdaBoost, *Gradient Boosted Decision Tree* (*GBDT*) models use gradient boosting in the loss function. In other words, in an iterative process, the weights of weak learners are adjusted to minimise the loss function. Several applications of GBDT models can be found in the literature for crash occurrence prediction and others. For instance, Lu et al. (2020) developed a gradient boosting crash occurrence prediction model for highway-rail grade crossing, which was compared to a simple decision tree model. Results showed that the gradient boosting model outperformed the simple decision tree model and revealed a nonlinear relationship between its determinants and crash likelihood.

A specific and more sophisticated implementation of gradient boosting decision trees is *Extreme Gradient Boosting* (*XGBoost*), which offers the advantage of computational efficiency. XGBoost's working is similar to GDBT as it learns from weak classifiers and adjusts/increases the weight of incorrectly classified samples, and the final prediction is obtained through ensembling (Chen and Guestrin, 2016). XGBoost models have gained recognition for their performance. For instance, Goswamy et al. (2023) developed an XGBoost model for investigating the factors affecting injury severity at pedestrian crossing locations with rectangular rapid flashing beacons and compared the model with random parameters discrete outcome models. This study found that the XGBoost model showed superior performance over its counterpart for injury severity prediction.

3.2.2. Neural network-based models

Neural network-based models are inspired by the working of biological neurons in the human brain. These models can be broadly classified into three categories as follows.

Artificial Neural Network (ANN) is the most fundamental type of neural network that mimics how human brain neurons work. In general, the architecture of ANN models consists of interconnected neurons and embodies three different components: input layer, hidden layer, and output layer. The input layer takes input data/features used for prediction, for which computations are performed in hidden layers, which are then fed to the output layer. Each feature in the input layer is multiplied with a corresponding weight assigned to a neuron unit. Then, the weightage sum of all input features plus a bias is calculated using a transfer function to minimise the difference between observed and predicted outputs. Using an activation function, the value of the transfer function is passed to determine whether the node should pass data to the output layer or not. In case of passing the value, the predicted value is calculated, which assists in determining prediction error. Based on the results, ANN will update its weights of nodes to minimise the prediction error iteratively until the lowest error is obtained and the convergence condition is satisfied — this method is known as back-propagation.

Optimising the hyperparameters of ANN models is crucial because they govern the learning and training process, and improperly selected hyperparameters will likely yield poor prediction performance. Common hyperparameters include (a) architecture parameters, such as the number of hidden layers, number of nodes per hidden layer, type of activation functions, and (b) training variables, such as learning rate, number of epochs, momentum, and batch size.

Since ANN models are the workhorse of the machine learning world as they can be applied to regression as well as classification problems, their application for crash prediction modelling is also diverse. Abdelwahab and Abdel-Aty (2001) modelled two-vehicle injury severities occurring at signalised intersections using ANN, which was reported to perform better than an ordered logit model.

Convolutional Neural Network (CNN) models, originally designed to work with images, create a network where each layer converts information from the previous layer into more complex information, which is fed to the next layer. Broadly, CNN models contain two blocks: feature learning block and classification block, whereby the former block is created from a number of convolution and pooling layers that are used to extract and learn features from the data, whereas the latter block classifies the learned/extracted features. All other characteristics of CNN are similar to ANN, e.g., training, activation function, hidden layers, etc. Our literature survey suggests that CNNs have been employed relatively less for crash prediction modelling, perhaps because these models work best with image data. However, there exist a few applications of CNN for crash prediction modelling. For instance, Hu et al. (2020) developed deep learning, CNN, and decision tree models using connected vehicle data. Fine-tuning of model hyperparameters was conducted using a Bayesian optimisation algorithm. Results indicated that the CNN model achieved relatively higher accuracy and was recommended as the classifier to predict the crash risk at intersections.

Recurrent Neural Network (RNN) models are specifically designed for modelling the serial dependency or recurrent pattern in time series data. Therefore, RNNs have feedback connections, enabling information dissemination between time intervals, unlike other types of neural network models. However, one problem in RNN models is the vanishing gradient, which limits their ability to memorise the information from distant intervals. To overcome this problem, Long Short-Term Memory (LSTM) models are considered, which possess the capabilities to capture long-term dependencies in sequential data. For this purpose, LSTM has multiple memory cells through which information is transmitted from one cell to another. Each cell receives input data from the current time step as well as information from previous time steps. To manage information transmission in the long-term, these cells contain three gates: forget gate, input gate, and output gate. The forget gate controls the information from the previous cell, which is meant to be forgotten in the current cell, whereas the input gate determines the information of the current input permitted to flow in the current cell. Finally, the output gate determines whether the information is released from the current cell. RNN and LSTM have been used for crash prediction modelling. For instance, Jiang et al. (2020) developed an LSTM-based modelling framework considering traffic data of different temporal resolutions for crash detection. Results revealed the satisfactory performance of the developed model (70.43% accuracy). Further, a transferability analysis was performed, indicating that the model developed for one freeway can be transferred to another with similar performance.

3.2.3. Other models

Support Vector Machine (*SVM*) models can handle complex non-linear classification phenomenon, whereby a higher *n*-dimensional space is created using a kernel method, which can linearise the relationship between output and input. The basic principle of SVM is separating the transformed data into different groups using an optimal *n*-1 dimensional hyperplane that maximises the distance from the hyperplane to the closest data points. Among several characteristics of SVM, striving to find global minima (because of convex optimisation) and being less prone to overfitting (because of the structural risk minimisation principle (Vapnik, 1999)) are the two most outstanding qualities.

SVM has been most widely used for all three aspects of crash prediction modelling. Yu and Abdel-Aty (2013) modelled real-time crash occurrence on a mountainous freeway and concluded that even a smaller sample size would enhance the SVM model's performance, variable selection methods are prerequisites to SVM's model training and testing, and explanatory variables showed identical effects on crash occurrence for both SVM and logistic models. Similar findings were reported in Yu and Abdel-Aty (2014).

Naïve Bayes models are based on Bayes' theorem and assume independence among explanatory variables. Given the vector of explanatory variables, the conditional probability distribution of a response variable (e.g., injury severity or crash occurrence) can be obtained using Bayes theorem. Naïve Bayes models are also frequently used for crash occurrence, crash frequency, and injury severity prediction purposes. For instance, Chen et al., (2016b) developed a decision table/Naïve Bayes hybrid classifier for injury severity analysis of rear-end crashes. The developed hybrid model performed better than the individual decision tree and Naïve Bayes models but showed poor performance compared to a multinomial logit-Bayesian network model.

Bayesian Network (*BN*) models consist of two parts: structure and parameters. The structure represents a network structure of explanatory variables defined by a directed acyclic graph, expressing dependencies and independencies among explanatory variables associated with each node. The parameters are associated with a conditional probability distribution associated with each node. Several advantages of BN models are reported (De Oña et al., 2011), including (a) no prior assumption required to develop a model, (b) providing a graphical representation of complex structures, and (c) causality between variables can be identified. Compared to other machine learning models, the application of BN models is still growing. Chen et al. (2015) proposed a multinomial logit model-based BN hybrid approach for injury severity analyses for rear-end crashes. The multinomial logit model identified significant contributing factors for rear-end injury severities, which were used to develop a BN model to explicitly formulate statistical associations between injury severity outcomes and explanatory attributes. The developed model shows a fair performance in predicting injury severity classes, with the maximum correct classification up to 65.7%. Notably, the model was not compared with any other competing model to justify its superiority.

K-Nearest Neighbour (*KNN*) models (also called instance-based learning with parameter *k* or lazy learning) are non-parametric and follow a distance-based metric to classify objects based on their proximate neighbour classes. In other words, KNN aims to classify the response variable (e.g., injury severity class or crash occurrence) based on the input features considering the majority vote of its top *k* nearest neighbouring features in the training dataset. KNN has been widely used for comparison purposes with other machine learning models like decision trees, artificial neural network, support vector machine, Naïve Bayes, and random forests. For instance, Iranitalab and Khattak (2017) compared statistical and machine learning models for injury severity prediction, including the multinomial logit model, KNN, support vector machine, and random forests. Results revealed the relatively poor performance of KNN models compared to other models. Similar (poor) performance was also observed by Theofilatos et al. (2019) when they compared deep learning models for crash occurrence prediction with KNN. Whilst KNN is mostly suitable for classification problems, it can also be applied to regression problems. However, the literature is devoid of any such application for crash prediction modelling.

Table A1 in appendix summarises the above machine learning models from different aspects and presents some characteristics and challenges of each model class.

4. A review of modelling intricacies

This section reviews several modelling intricacies related to machine learning models, which are by and large ignored in previous review studies. These intricacies include but are not limited to hyperparameter selection, data source, sample size, selection of performance metrics, and so on, which are explained in the ensuing subsections.

4.1. Selection of hyperparameters

Hyperparameter tuning (or selection) is the key to leveraging the full potential of a machine learning model in providing the highest predictive power. Whilst some studies have selected hyperparameters for their models using random searches (Ijaz et al., 2021, Das et al., 2020), many studies have used either grid search (Lu et al., 2020, Li et al., 2012) or Bayesian optimisation (Dong et al., 2022, Yang et al., 2022). Random search (or providing a user-specified range) has been used for determining appropriate model hyperparameters, which-for unknown reasons-has been adopted by several studies. For instance, Ijaz et al. (2021) used a range of hyperparameters for machine learning models to predict injury severity crashes involving three-wheeled motorised rickshaws. Results revealed that selecting appropriate hyperparameters from the specified range yielded the highest performance. Grid search is often performed to determine the optimal hyperparameters for a given model. This search is performed for a user-defined range, which can be exhaustive or limited. However, an issue associated with exhaustive grid search is the computational time required to run all the possible combinations of hyperparameters. Using the grid search for random forest model developed for injury severity analysis, it was reported that the model performed slightly better than an ordered probit model (Li et al., 2012).

Unlike grid search, Bayesian optimisation utilises past evaluation results that were used to develop a probabilistic model for mapping hyperparameters to a probability of a score on the objective function. In general, Bayesian optimisation is an effective method for selecting hyperparameters as it chooses the next hyperparameters in *an informed manner*. The basic principle is to meticulously select the next hyperparameters to minimise the objective function. Considering hyperparameters that appear the best from past results, Bayesian optimisation envisages finding a better model setting than random search in fewer iterations. In order to predict pedestrian fatalities, Yang et al. (2022) used Bayesian optimisation for hyperparameters tuning for models, such as support vector machines, ensemble decision trees, and *k*-nearest neighbours (KNN). Although KNN performed poorly compared to other models in predicting pedestrian fatalities, the performance gain for KNN due to Bayesian optimisation was the highest among all three machine learning models.

4.2. Data source

For developing a machine learning model, data play a pivotal role. To this end, datasets used by previous studies can be broadly classified into three groups: police-reported data (or data provided by different departments of transportation), driving simulator data, and probe data.

Police-reported data is the most commonly used data source for crash models, whereby information about a crash is recorded and transcribed by local police. These crash records are supplemented with traffic conditions, geometric characteristics of roadways, and environmental factors that assist in generating non-crash events for developing a real-time crash prediction model. These additional data are obtained through road transport authorities (e.g., departments managing loop detectors on highways), metrological departments, and geographical drawings of roadway infrastructure. Using a Bayesian Belief Net model, Hossain and Muromachi (2012) used loop detector data to supplement crash records for real-time crash occurrence prediction. Results revealed that using an average threshold value, the model can successfully classify 66 % of future crashes with a false alarm rate of less than 20%.

Whilst police-reported data have several well-known issues (presented in Section 5), one of the main issues is the lack of behavioural factors associated with the crash occurrence or injury severity. For behavioural research, driving simulator experiments are utilised. The driving simulator data includes more granular information about driving and road user trajectories. For instance, Abou Elassad et al., (2020b) conducted a driving simulator study for predicting real-time

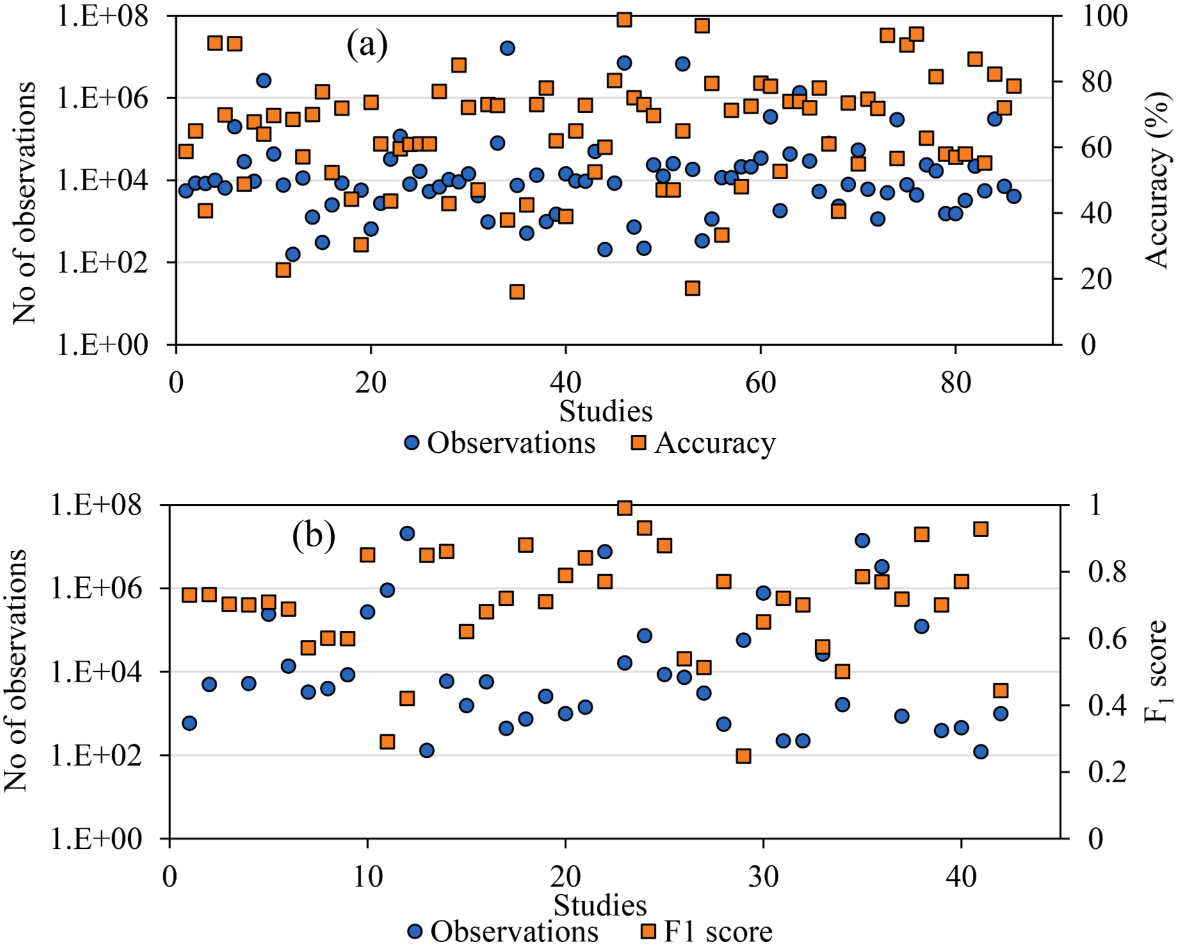

Fig. 4. Sample size used and corresponding model accuracy for (a) injury severity models and (b) crash occurrence models.

crash occurrence. The dataset contained physiological, driving behaviour, and weather information, which was fused together to better understand real-time crash occurrence. It was reported that using this dataset combined with a boosting mechanism (AdaBoost) yielded high prediction results for real-time crash occurrence.

Another source of collecting real-time data is probe data generated by connected vehicles, autonomous vehicles, or any instrumented vehicles (i.e., naturalistic driving studies). These datasets are larger than those collected by driving simulator studies and police-reported records. Whilst these datasets require significant pre-processing, they provide unprecedented opportunities to study crash mechanisms microscopically. For instance, Zhang and Abdel-Aty (2022) used connected vehicle data provided by Wejo Data Services, Inc., with a data frequency of 3 s. This dataset was fused with loop detector data obtained from Central Florida Expressway Authority and the Florida Turnpike, containing spot speed, lane occupancy, and volume information every 30 s, whereas crash data were obtained from the Signal Four Analytics (S4A) database. Using a bidirectional long short-term memory (LSTM) model with two convolutional layers to predict real-time crash potential on freeways, this study reported that the bidirectional LSTM model outperformed other competing models in predicting real-time crash occurrence.

4.3. Sample size, performance metrics, and model performance

Sample size plays a critical role in machine learning models, and studies have used a wide range of sample sizes for their models, as evident in Fig. 4. Note that this figure is developed with studies that have used similar performance metrics, and as such, 86 and 42 studies are used injury severity and crash occurrence, respectively. Note that this plot is not developed for crash frequency because of inconsistent metrics used for analysis. Notably, the number of observations as low as 122 (Li et al., 2008) and as high as 20,534,532 (Morris and Yang, 2021) were used for crash frequency and occurrence models, respectively. Whilst most studies for crash frequency have used a smaller sample size, the largest sample size is predominantly used in crash occurrence models, whereby non-crash events are considerably higher than crash events.

A general notion for machine learning models is that the model is likely to yield better performance with an increased sample size. Although a cross-study comparison would be difficult because of confounding factors, Fig. 4 (a) shows that models developed with a higher number of observations do not necessarily attain the maximum accuracy. For instance, Morris and Yang (2021) found the varying performance of the extreme gradient boosting model with different resampling techniques, whereby the accuracy varies from 3% for single vehicle crashes to 82% for non-crashes, whilst the weightage average for all crashes varies between 57% and 67% (note that this study has 20,534,532 observations).

Machine learning models are assessed using a variety of performance metrics, with different indicators for each type of crash modelling aspect. For crash occurrence and injury severity prediction models, most models have been evaluated using accuracy, sensitivity, specificity, precision, F_1 score, an area under the receiver operating characteristic (ROC) curve, Cohen's Kappa, and geometric mean (G-mean). For instance, Zhang et al., (2022a) used accuracy, Cohen's Kappa, F_1 score, and area under the receiver operating curve (AUC-ROC) for analysing injury severity in single and multiple-vehicle crashes using Naïve Bayes, *k*-nearest neighbour, binary logistic regression, and extreme gradient boosting (XGBoost). Results revealed a better performance of the

Table 3
A representative list of explanatory variables used in crash modelling.

Crash aspect	Explanatory variables (or features)
Occurrence	*Traffic operational characteristics*: Network speed, lateral acceleration, longitudinal acceleration, yaw rate, pedal position, flow, occupancy, congestion index, average speed, standard deviation of speed, traffic volume, and several variants of speed (80th percentile, standard deviation, coefficient of variation), green time ratio for left and through traffic
	Environmental: Weather conditions, humidity, precipitation, temperature, windspeed
Frequency	*Traffic operational characteristics:* Traffic volume, average offset distance to roadside fixed objects, average annual daily traffic
	Geometric: Segment length, horizontal alignment, shoulder width, segment length, horizontal curve radius, absolute value of deflection angle, vertical curve radius, slope gradient, slope length.
	Human factors: Fatigued driving and gender, density of commercial driveways, density of industrial driveways, density of residential driveways, vehicle miles travelled, speeding, disobeying traffic rules
	Crash characteristics: Year, season, vehicles involved in a crash
Injury severity	*Traffic operational characteristics*: Disobeying traffic rules, pedestrian movement, sight distance, posted speed limit on mainline, difference of speed limit between mainline and exit ramps
	Environmental: Weather conditions, lightning
	Geometric: Segment length, type of shoulder, road surface condition, number of lanes in mainline road, at on-ramp, at off-ramp, length of deceleration lanes, mainline annual daily traffic (ADT), exit ramp ADT, safety barriers, pavement markings
	Human factors: Gender, age, seat belt use, drugs use
	Crash characteristics: Crash type, time, day of the week

XGBoost model compared to other models for injury severity prediction.

For crash frequency models, several metrics are used, including mean absolute error, mean squared prediction error, mean absolute deviation, average squared error, standard deviation of error, mean absolute percentage error, normalised mean squared error, explained variance, *R*-square, and a 45-degree validation plot. Bao et al. (2019), for instance, used mean squared error, mean absolute error, and mean absolute percentage error to evaluate their spatiotemporal deep learning model for citywide short-term crash risk prediction with multi-source data. Using these error metrics, this study reported better performance of the deep learning model for citywide short-term crash risk prediction by leveraging multi-source datasets.

4.4. Explanatory variables, variable importance, and selection

Using an adequate set of explanatory variables (or features in machine learning — note that they are interchangeably used in this review paper) is essential to fully leverage the capabilities of machine learning models. Recognising the complex nature of crashes, several factors are used, which can be broadly classified into five groups: roadway geometry, vehicle operational characteristics, crash characteristics, environmental, and human factors, which are summarised in Table 3.

The selection of an appropriate subset of explanatory variables is important to train a machine learning model. To this end, variable/feature importance and selection become paramount and should be performed in data preparation. Our review suggests twofold findings in this direction. First, about half of the studies have not assessed variable importance for their models. Second, about three-quarters of the studies have not employed any variable selection technique. Some decision tree-based methods can calculate the importance of explanatory variables, e.g., classification and regression trees (Basso et al., 2018, Basso et al., 2020). Chen et al., (2016a) employed classification and regression trees to obtain variable importance ranking and perform the variable selection for injury severity modelling. This study found that four variables (out of 22) did not contribute to injury severity prediction and were left out of the parsimonious model. Apart from this technique, several other techniques have been used, including but not limited to chi-square (Ghandour et al., 2020), random forest regression (Ahmad et al., 2022), generalised linear models (Ji and Levinson, 2020), SHapley Additive exPlanations (Wei et al., 2022), Shapiro–Wilk (Sinha et al., 2021), and many others.

4.5. Overfitting

Overfitting is a common issue in machine learning models, whereby models provide high prediction accuracy for training data and significantly lower performance on testing/validation data. One of the consequences of an overfitted model is the lack of generalisation capabilities to new datasets, which limits a model's broad applicability. Overfitting can occur for several reasons. First, a small training sample, which fails to present an overall picture of underlying patterns in the entire data, leads to significant discrepancies between training and testing dataset predictions. Second, noise in the data can contribute to overfitting as the model will falsely assume noise as data patterns. Third, training a model for a sufficiently long time, even after model convergence, can also lead to an overfitted model. Finally, if the model complexity is unnecessarily high for a simpler dataset, this complexity may cause learning the noise within training data and lead to overfitting. To overcome overfitting, a common way is to use *k*-fold validation, whereby the entire dataset is divided into *k* equal sizes (or folds) and each fold is used for training and testing, thereby ensuring that the model has been rigorously tested with all patterns available in the data. To this end, there is a great need to further use overfitting techniques to tackle this problem (see Section 5 for more details).

4.6. Data imbalance

Data (or class) imbalance is a serious issue in crash occurrence and injury severity datasets, which refers to a significantly higher number of observations of one class than the other class(es). This imbalance leads to a biased learning model exhibiting a high accuracy in predicting the majority class and a poor prediction performance for the minority class. For crash occurrence prediction models, the degree of data (or class) imbalance is determined by how non-crash events are extracted/obtained. For this purpose, past studies have either used (matched) control-to-case ratio (Sun et al., 2014) or a time slice before a crash is considered a non-crash event (e.g., 12 s before a crash (Abou Elassad et al., 2020b)). Sun et al. (2014) developed a support vector machine model for real-time crash occurrence prediction on expressways. Their data contained 90 crashes, and non-crashes were obtained using a matched case-control design with a ratio of 1:10. This study did not compare their model performance with varying case-control ratios. Contrastingly, using a time window of 5 mins before a crash, Zhang and Abdel-Aty (2022) developed a bidirectional long short-term memory model for predicting real-time crash occurrence on freeways using connected vehicle data. Synthetic minority oversampling technique was used to handle class imbalance. This study also tested four oversampling ratios, and the 1:4 ratio provided the highest model prediction performance.

Several techniques have been tested to account for class imbalance, such as random over-sampling, synthetic minority over-sampling, adaptive synthetic sampling, Wasserstein Generative Adversarial Network, and others (Man et al., 2022b, Morris and Yang, 2021, Parsa et al., 2019). Resampling methods have been found to enhance model performance, and the adaptive synthetic sampling approach performed relatively better than others (Morris and Yang, 2021). However, the theoretical underpinning behind these techniques and their relation to practical situations is relatively weak.

4.7. Model validation

Model validation is a critical step in ensuring the performance of machine learning models and assessing their overfitting/generalisation ability. Predominantly two approaches are used for model validation,

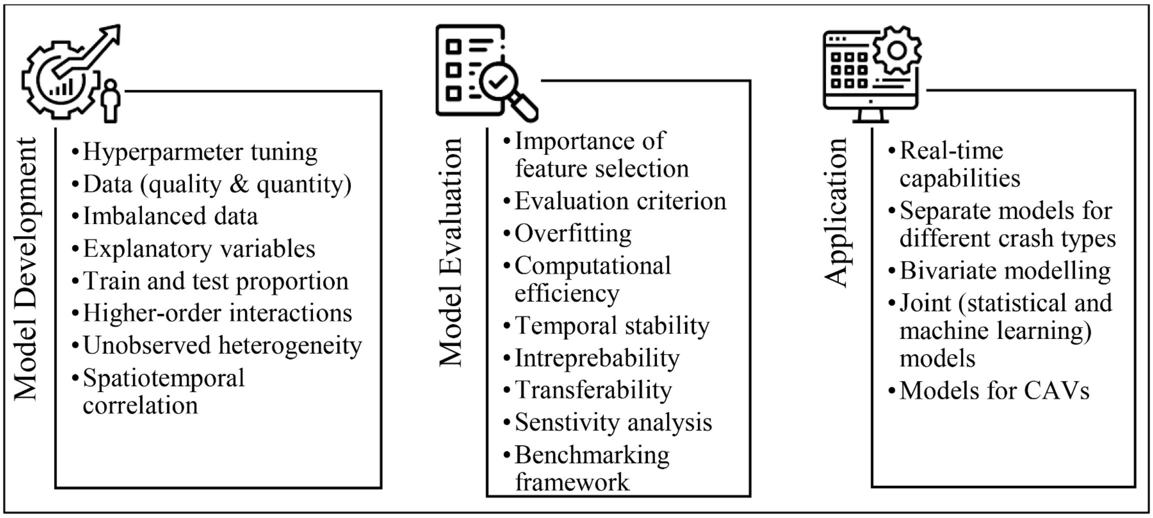

Fig. 5. Research needs for developing machine learning-based crash models.

namely *k*-fold validation and simple train-test split based on hold-out sampling. In a *k*-fold validation approach, the entire dataset is divided in small and unique *k* folds, ensuring that a model is trained and testing on entire dataset in small chunk without overfitting. Our review suggests 5-, 10-, and 20-fold validations are commonly used, with 10-fold being more prominent than others. For instance, Ijaz et al. (2021) compared three machine learning methods for predicting injury severities of three-wheeled motorised rickshaw crashes at signalised intersections. Their study used 10-fold validation to assess their model performance, and results indicated decent performance of the decision jungle model, outperforming other competing models.

Contrastingly, a train-test splitting approach generally assesses the accuracy of a model on out-of-sample test data, which is generally considered to prevent overfitting and improve transferability (Browne, 2000, Picard and Cook, 1984). Within cross-validation, the hold-out method is prevalent, whereby the entire dataset is randomly split into training and testing sets (e.g., 70 % for training and 30 % for validation). For instance, for crash frequency modelling, Zhou et al. (2022) integrated machine learning and statistical models to gain insights into the relationships between explanatory variables and crash frequencies and then used machine learning models to improve the predictions obtained from statistical models. This study used crash frequency data, with a 70–30 training–testing classification ratio and found that their integrated approach significantly improves the predictive performance and goodness-of-the-fit of statistical models without adversely impacting their interpretability.

4.8. *Interpretability*

Machine learning models have long been called black boxes due to their poor interpretability. However, recent advancements in explainable artificial intelligence have overcome this major limitation of machine learning models. Our review suggests that almost 80% of studies have not employed any interpretability technique to explain the relationship of model outcome with its determinants. The remaining studies have applied several techniques for uncovering the relationship, including Local Interpretable Model-agnostic Explanations, Local Sensitivity Analysis, Partial Dependence Plots, Global Sensitivity Analysis, and SHapley Additive exPlanations. A detailed mathematical explanation of these methods can be seen in Wen et al. (2022), whereas a representative application of these methods is explained here with their general working mechanism.

Local Interpretable Model-agnostic Explanations (LIME) train a local alternate model for approximation and then use it for machine learning model predictions (Ribeiro et al., 2016). Particularly, LIME generates model outcomes as a function of perturbating original observations that are fed into a trained model, leading to a new dataset where perturbed observations are paired against their corresponding model outcomes. An interpretable model, e.g., linear regression, is then trained on this new dataset, whereby new observations are weighted to the original observations based on their proximity. As an outcome, the original observation can be explained with the alternate model. Arteaga et al. (2020) developed an injury severity model by analysing traffic crash narratives to identify factors associated with high injury severity levels. Using LIME, this study identified potential causality factors for injury severities whilst providing interpretability as required by traffic safety analysts. Results revealed the ranking of the significant factors, which were found to be comparable with tabular data and classical regression analyses.

Another interpretability technique is partial dependence plots (PDP), which uncover the marginal effects of one or two explanatory variables on the model outcome by considering the average effects of all other explanatory variables (Schlögl et al., 2019). PDP can be useful for interpreting how one variable affects the model outcome and can also provide insights into the selected subset of variables' effect on the model outcome. For instance, Ding et al. (2018) developed a pedestrian crash model using a multiple additive Poisson regression trees method and applied PDP to understand the non-linear effects of the built environment on pedestrian crash frequency. Results revealed the presence of non-linear effects of certain variables, including road network characteristics, street elements, land use patterns, and traffic demand, which suggested the local road authorities should adopt geo-spatial differentiated policies to establish a safe walking environment.

Global sensitivity analysis (GSA) is another widely used method for model interpretation. Contrasting to partial dependence plots, GSA determines the subset of explanatory variables from a list of explanatory variables. The change in the model outcome is a function of averaging the changes in (subset) explanatory variables expressed in probability distribution (Saltelli et al., 1999). Using customised Monte Carlo methods, the variance in the outcome variable is decomposed and linked to either an individual explanatory variable or a set of explanatory variables (Jiang et al., 2022). Using this method, Jiang et al. (2022) developed a crash severity model using ensemble methods (i.e., AdaBoost and Gradient Boosting) and applied GSA to determine the individual and joint impacts of explanatory variables on crash severity outcome. Results showed that the most influential factors were the vertical curve, seat belt use, crash type, road characteristics, and truck percentage. Further, using a simulation approach, this study demonstrated that an explanatory variable may have a varying impact on crash severity when it changes within specific ranges.

SHapley Additive exPlanations (SHAP) is a technique based on game theory for understanding the players' contribution to the success of a collaborative game (Lundberg and Lee, 2017). SHAP is theoretically sound and can be applied to any machine learning model to provide local and global explanations. The SHAP analysis is often employed to explore the importance and impact of each explanatory variable on the model prediction performance (Nikolaou et al., 2023). Parsa et al. (2020), for instance, applied SHAP to understand real-time characteristics of crash detection. Using SHAP, feature importance was demonstrated, revealing speed and traffic volume after a crash at the upstream location as the two most important features. Further, using a feature dependence analysis characteristic of SHAP, the interaction effect of two variables on crash occurrence was also discussed.

To summarise, local interpretation methods like LIME, LSA, and PDP tend to make inherent assumptions about explanatory variables being independent of each other, which may not be valid in reality. As such, the interpretation offered by these methods may lead to biased safety estimates (Wen et al., 2021). This limitation appears to be overcome by SHAP, which can calculate total, main, and interaction effects.

5. Research needs in developing machine learning-based crash prediction models

This section summarises general challenges and future research directions based on the thorough literature survey presented in Sections 3 and 4. More specifically, research needs that apply to crash prediction modelling using machine learning are mapped into three groups, as shown in Fig. 5. These needs are discussed in the ensuing subsections.

5.1. Model development

Developing a machine learning model considering many factors involved during its development is not straightforward. This section highlights some critical aspects of model development and related issues that need to be meticulously addressed.

5.1.1. Hyperparameter tuning

Hyperparameter tuning is critical in developing a robust machine learning model. In essence, the model performance is a function of hyperparameters (or their tuning). As such, an optimal set of hyperparameters should be used. However, selecting optimal hyperparameters is difficult and highly depends on the hyperparameter tuning technique used. Whilst random search requires testing different combinations of parameters, prior knowledge from past studies about hyperparameters can be leveraged to expedite the process. Contrastingly, grid search is a comprehensive technique that tests all the possible combinations and provides the optimal hyperparameters. However, the time taken by grid search can be too long if the number of parameters involved is large and their ranges are exhaustive. One way to address this problem is to employ a halving technique that searches over a specified list of hyperparameters using a successive halving approach. Particularly, all the possible combinations are tested on a small sample, through which the best possible combinations are iteratively selected and tested on larger samples, thereby reducing the computational time. Bayesian optimisation is another technique that relies on Bayes grid search and considers the past prediction results when selecting new hyperparameters, which are likely to provide better prediction results in a shorter time.

Whilst all these methods have pros and cons, there is no easy answer to the question: which technique to select?. On the one hand, the motivation is to select the optimal parameters that yield the highest possible predictive accuracy, but on the other hand, it is also required that such selection should be computationally efficient because of the reasons mentioned in Section 5.2.4. To overcome this problem, a rigorous comparison of these hyperparameter tuning methods should be performed, and subsequently, their effects on model performance should be quantified as a function of computational efficiency.

5.1.2. Data quality and quantity

Data are an integral part of model development that has significant implications on model performance. Both data quality and quantity are crucial for modelling. The quality of data highly depends on how it was measured or recorded. Several known issues exist when data are recorded by humans (such as police-reported data); some of them are underreporting, low sample means, limited behavioural information, and omitted variable bias (Ali et al., 2023b). Understandably, these issues represent the typical nature of crash data, but it is desirable to investigate how these issues impact the model performance. To overcome these issues, a possible direction could be data fusion, where multiple data sources could be fused together to provide richer information (Ali et al., 2023a). For instance, trajectory data obtained from video recordings could be geospatially (or geo-temporally) complemented with police-recorded data, providing information about driving behavioural factors. Using drones and LIDARs also provides a new way of exploring road safety with rich behavioural information. Similarly, using traffic conflicts instead of crashes is a new paradigm that has shown considerable promise in predicting crashes using non-crash data (Hussain et al., 2022). To this end, machine learning is clearly needed to link crashes with conflicts, which has received little attention in the literature (Ali et al., 2023a).

The second aspect of data relates to quantity. A general hypothesis about machine learning models persists that increasing the sample size will lead to better prediction power. This hypothesis could be true if the increased sample provides new information to the model. Contrastingly, if no new trends or patterns are found in newer observations, such data may become redundant instead of increasing model predictive power (Chen et al., 2016b). In the literature on crash prediction modelling, very little has been explored on the effects of sample size for machine learning models. A sensitivity analysis should be performed to obtain the optimal sample size for a given problem (see Zhang and Abdel-Aty (2022), who compared four sample sizes and found an intermediate sample size to be the best).

Another aspect of data is the significant underreporting of low-severity crash types. For instance, property damage-only crashes are often not reported to avoid potential traffic fines and increase insurance premiums (Blincoe et al., 2002). This issue is exacerbated in some countries' databases when property damage-only crashes are neither counted nor included in the crash classification system (Imprialou and Quddus, 2019). Past research reported varying underreporting rates of crashes across countries (Janstrup et al., 2016, Watson et al., 2015, Yannis et al., 2014), which is a function of driver's age, crash severity level, road type, and so on (Abay, 2015, Amoros et al., 2006). Most fatal and severe crashes are documented in police-reported data, whereas two-thirds of minor and property damage-only crashes remain unreported (Imprialou and Quddus, 2019). Data fusion could be performed to overcome this underreporting issue, whereby data from other sources could be leveraged. For instance, data from road crash databases, hospital admitted patients records, emergency department information systems, and injury surveillance units could be combined to determine the underreporting of a crash, which can be used to resample during the model development.

When combining data from multiple sources, two issues emerge: greater consumption of network resources and the possibility of exposing sensitive information. To overcome these issues, federated learning can be applied, leveraging the computational power of the agents collecting data, which can be used for the model learning process with locally available datasets.

Apart from adopting data fusion to improve data quality and quantity, preliminary data analyses (e.g., descriptive statistics, sample size, and outliers) should be performed to understand the data limitations.

Along the same lines, the robustness of machine learning models should be tested against well-known issues like omitted variable bias. Further, future research should develop intelligent crash reporting systems, which become more relevant in the era of connected and autonomous vehicles. These systems can utilise data from in-vehicle sensors, LiDARs, and cameras that will automatically record trajectories of all contributing vehicles without human intervention. Evidence of some intelligent reporting systems already in place are Traffic and Criminal Software (TRACS) in the US, Collision Recording and Sharing (CRASH) in the UK, PRC and ReGIS in Italy (Montella et al., 2019, DfT, 2011, Ogle, 2007). Finally, the analysis of the impact of data inaccuracies (e.g., inaccurate crash location, time, severity, and determinant) needs to be quantified, and post-processing methods (e.g., crash mapping) should be developed for improving data quality and addressing crash data inaccuracies. Apart from data fusion of numeric data as mentioned above, integrating multimodal data is another possibility, whereby text, images, and sensor data could be fused using multimodal deep learning techniques for more robust predictions.

Finally, as the crash database provides information about the crash frequency and resulting injury severity, there is a clear need to link crash frequencies with injury severities, providing in-depth insights into the likelihood and frequent nature of different injury crashes. With high-dimensional data already being used, the use of advanced modelling methodologies becomes imminent, e.g., Generative Adversarial Networks or Variational Autoencoders. These methods offer nuanced insights by modelling the underlying data distribution more effectively, with specialised architectures being particularly useful for capturing spatial–temporal patterns, which are often critical in predicting different crash aspects.

With fundamentally different nature of the three aspects of crash evident in data, the use of specialised techniques for a given crash aspect merits an application. For instance, the complexity of predicting injury severity could benefit from more specialised classification techniques, e.g., ordinal regression neural networks that are suitable for modelling the ordinal nature of injury severity categories more effectively.

5.1.3. *Treating imbalanced data*

Data (or class imbalance) is a major issue in classification-based machine learning models (in this study context, i.e., crash occurrence and injury severity prediction models). Our review found that over 70% of crash occurrence and injury severity models did not account for class imbalance. As such, their findings are likely to be biased towards the dominant class, e.g., non-crash events in crash occurrence prediction models. In this context, past studies used methods like random sampling, (matched) case-control design, and a fixed time window to account for data imbalance in crash occurrence models. Whilst random sampling is easier to implement and may reduce analyst bias when multiple trials are performed, it does not consider any prior knowledge available for non-crash events. Meanwhile, selecting non-crash events against crashes naturally fits into the framework of the case-control design. One important aspect of this design is determining how many controls (non-crashes) should be selected against one case (crash). Inconsistent selection of controls has been found in the literature whereby 1:2, 1:3, 1:4, and several other case-control ratios are selected (Ma et al., 2022, Cai et al., 2020, Theofilatos et al., 2019). Future research should consider determining the optimal case-control ratio for real-time crash occurrence prediction. Finally, using a slice of time window prior to a crash has also received significant attention in recent studies, whereby different time windows are used, e.g., 12 s to 5 mins before a crash (Basso et al., 2021, Abou Elassad et al., 2020a, Abou Elassad et al., 2020b). However, inconsistencies are found in selecting the time slice, partly due to the lack of theoretical knowledge and a systematic approach to selecting different time windows. Therefore, different time windows should be set and compared to decide the most suitable one for real-time safety analysis. Along this line, the application of zero-shot and few-shot learning appears relevant, which learns and makes predictions from a few or even zero labelled data that is often prevalent in real-time crash occurrence modelling.

Two approaches are used for handling class imbalance in injury severity models: undersampling and oversampling. Whilst the former approach aims to balance the ratio of classes by undersampling from the majority class, the latter approach oversamples the minority class to match the proportion of the majority class. Clearly, oversampling prevents data loss compared to undersampling, where samples are cut down to balance the class proportion. The Synthetic Minority Oversampling Technique (SMOTE), presented by Chawla et al. (2002), is frequently used for handling class imbalance in safety literature. However, most studies do not consider comparing SMOTE's performance with other oversampling techniques as SMOTE is criticised for being overly generalised and prone to overfitting (Mease et al., 2007).

Several other sophisticated techniques, such as variational autoencoder (Islam et al., 2021), generative and adversarial network (GAN) (Cai et al., 2020), and Wasserstein GAN (Man et al., 2022a), are specifically designed to produce highly realistic and diversified samples to account for class imbalance. Transformer-based models, which have shown great promise in various classification tasks, could also be explored for their ability to handle imbalanced datasets and high-dimensional feature spaces for injury severity models. Future research should holistically compare these sophisticated techniques and provide guidelines on which technique should be preferred for class imbalance. Along this line, it is worth investigating whether one technique (e.g., GAN) works better for different crash types, which can be assessed by comparing different crash types (e.g., rear-end crashes versus sideswipe). Another important research question appears to be unanswered: what is the relationship between the degree of data imbalance and model performance?. Answering this question will assist researchers in interpreting their results with caution when class imbalance is not treated. Finally, do these sampling techniques perform similarly for crash occurrence and injury severity models, or are they application-specific?.

5.1.4. *Explanatory variables*

Explanatory variables (commonly called input features in machine learning) play a pivotal role in model performance. Whilst different variables are used in different machine learning-based crash prediction models, driving behavioural factors that provide in-depth insights into crash mechanisms (Ali et al., 2022) are largely missing from these models, hindering our understanding of what driving behaviour is more likely to lead to a crash and which are the most significant factors. The omission of driving behavioural factors from crash models can be partly attributed to data sources used in earlier studies, i.e., sensors and loop detectors. To obtain driving behavioural factors, advanced data collection systems, such as video data collection, LiDARs, and drones, can be used to generate trajectories, which can provide insights into driving behaviour, and subsequently can be used as input into crash prediction models.

Finally, our review suggests that most studies developed machine learning models considering different crash types (Mokhtarimousavi et al., 2019, Mokhtarimousavi et al., 2021). Factors affecting crash types significantly differ from each other, e.g., factors affecting rear-end crashes are driver's response time, distance to the leader, relative speed, and others, whereas the factors affecting lane change crashes are lag and lead gaps in the target lane, relative speeds in the target lane, and others. Therefore, models considering all crash types together will provide biased predictions since the set of explanatory variables can explain one type of crash better than the other. To overcome this issue, a rigorous comparison of individual models and combined models merits an investigation.

5.1.5. *Uniformity in training and testing data proportion*

From our review, it was found that at least 23 different training–testing data proportions are used, with some generally used proportions like 70–30 or 80–20 to very unusual ones like 49–51, 63–37, and 95–5 (note that these numbers reflect percentages). Even inconsistencies have been found during the model development regarding whether data should be split into training and testing proportions or training, validation, and testing proportions. These inconsistencies hinder comparing models across different studies, limiting advancements in future model developments.

In essence, the training dataset represents the largest chunk of the original dataset used to train the model, i.e., understanding the patterns in the data. The validation dataset assists in hyperparameter tuning and selecting the appropriate model. Finally, the testing dataset is used to assess the performance of a fine-tuned trained machine learning model. Although there is no guideline on whether to use training–testing proportions or train-validation-testing proportions, it is suggested to use the latter segmentation of the data because a dedicated proportion of the dataset will be used to fine-tune the model. Further, some important research questions need to be answered. (1) Do different training–testing proportions affect the model performance?. If so, there is a need to quantify the change in model performance as a function of training–testing proportion. (2) Does the segmentation of the original dataset into training–testing proportions or training-validation-testing proportions have any effect on model performance?. To overcome this problem, one can select a training–testing proportion (e.g., 70–30), divide it into training, validation, and testing (e.g., 70–15-15), and assess the model performance. We hypothesise that model performance will vary when the testing dataset is divided into validation and testing; however, this hypothesis needs to be rigorously tested.

5.1.6. Considering higher-order interactions

Given the complexity of crash data and the multitude of factors involved in a crash, it has been mentioned that explanatory variables' main effects and interaction effects should be considered during the model development (Ali et al., 2021). Whilst main effects are relatively straightforward to include in a model, interaction effects are difficult to determine, which tend to grow geometrically and exponentially, respectively, with the number of ordinal and nominal variables (Haque et al., 2016). Machine learning models are notorious for ignoring causal effects, and it appears that most studies only considered main effects. Not considering interaction effects may have a significant impact on model performance.

To this end, a systematic approach is required to determine and test potential interaction effects in a machine learning model. For instance, Chi-square Automated Interaction Detection (CHAID) can be used (Ramotowski and Fitzgerald, 2020), which has shown the potential to extract all the possible higher-order interactions. Similarly, feature subset selection algorithms based on association rule mining can also be used, leveraging their structure freedom and global search capability (Hu et al., 2022).

5.1.7. Accounting unobserved heterogeneity

The data scarcity regarding all the factors contributing to a crash leads to an unobserved heterogeneity issue, which has significant implications for model inferences and specifications (Mannering et al., 2016). Several advanced statistical and econometric methods exist, such as random parameters and latent class models, which address this issue. However, machine learning models largely ignore this issue, assuming that a crash (aspect) is solely related to variables observed in the data. However, in reality, crash databases do not contain information about all variables, and as such, this strict assumption leads to omitted variable bias during model development. With significant advancements being made in statistical and econometric models to capture unobserved heterogeneity, there is a clear need to apply those learnings in advancing machine learning models to capture unobserved heterogeneity. For instance, akin to a random parameters model whereby parameters are allowed to vary rather than fixed, machine learning models could also assign different weights instead of fixed weights for predictions. Similarly, recent developments in machine learning models like Neural Graphical Models (Shrivastava and Chajewska, 2022) can be used that learns to represent the probability function over the domain using a deep neural networks.

5.1.8. Capturing spatiotemporal correlation

As crashes exhibit spatiotemporal correlation (Ihueze and Onwurah, 2018, Quddus, 2008), capturing both temporal and spatial correlations becomes paramount during model development. Whilst most machine learning models appear to ignore such correlation (i.e., crash occurrence and frequency prediction models), some models use temporal indicators, such as weather, week, and time of day, to account for temporal correlation in injury severity models. However, an issue with such models is that they consider crashes as independent observations, implying that spatiotemporal correlations do not exist, which can significantly impact their model performance (Wen et al., 2021). To this end, advanced machine learning models, such as graph convolutional networks, may be used, leveraging the topological characteristics of graph convolutional networks to capture spatial and temporal dependencies among contributory factors. Further, recurrent neural network or long-short term memory neural network models that possess the inherent capability of capturing correlation can be used. Along this direction, another related issue that has been overlooked especially in crash frequency models is predicting crash frequency trends for future years. Whilst significant work has been done in validating crash frequencies, the knowledge obtained from these models needs to be utilised to better understand how crash patterns evolve over the years and how different interventions could reduce/increase crash frequency.

5.2. Model evaluation

Evaluating the performance of a machine learning model is essential to make accurate predictions. This section provides insights into several intricacies that need be considered during model evaluation.

5.2.1. Importance of explanatory variable (or feature) selection

When a machine learning model is developed, it is essential to assess whether all the explanatory variables considered in the model contribute to the model prediction. Note that the issue of variable selection applies to all three crash aspects (i.e., crash occurrence, injury severity, and frequency). A model may result in poor predictions because of insignificant explanatory variables present in a model (Chen et al., 2016b). Therefore, a parsimonious model may be derived based on sequentially eliminating all the insignificant variables that do not contribute to model predictions. Several techniques have been used in the literature to obtain the importance of explanatory variables, such as classification and regression trees, random forest, and principal component analysis. Given the inherent difference in these techniques, it is worthwhile to investigate whether explanatory variable selection by different methods impacts model performance. Answering this question will provide insights into whether a particular technique should be used or the decision should be left to the analyst's discretion.

5.2.2. Evaluation criterion

Machine learning models are assessed using different performance metrics, such as F_1 score, for classification models (crash occurrence and injury severity prediction) and mean absolute error for regression models (crash frequency prediction). Our review finds inconsistent definitions of outcome variables for injury severity models, with two, three or more injury classes used for modelling and assessed using several metrics (see Section 4.4 for more details). Machine learning models (especially tree-based) are found to be biased towards having more levels. Therefore, using performance metrics like accuracy will likely provide misleading and biased estimates. To this end, metrics considering predictions for all classes, such as F_1 score and area under the receiver operating curve, should be considered. In fact, a two-step procedure can be adopted: (a) develop a base model with two categories, and (b) investigate the model performance (gain or loss) using appropriate classification metrics (e.g., micro and macro average receiver operating curve) by adding more injury classes, which should also be theoretically justified.

Conventionally, the performance of machine learning-based crash prediction models is assessed deterministically based on point estimates (e.g., one F_1 score) to justify the selection of one model over competing models. Instead, statistical analysis should be performed to compare whether there is a statistically significant difference in model predictions or otherwise. Such analysis will help decide between two (or more) models that may have similar performance scores, but their spread is different, with wider spread reflecting uncertainty in model predictions that provide a theoretical justification to select one model over others.

Past studies compared machine learning models with simple conventional statistical models, such as logistic regression and multinomial logit models, with a few exceptions. However, there is a clear need to compare machine learning models with recent advancements in statistical and econometric models, such as random parameters with heterogeneity in mean or variance and correlated random parameters heterogeneity in mean or variance. With some studies have already highlighted the better performance of statistical models compared to machine learning (e.g., Yu and Abdel-Aty (2014) showed that a random parameter outperformed a support vector machine model), there is a clear need to holistically compare different machine learning models with variants of advanced statistical and econometric models.

5.2.3. *Overfitting*

Machine learning models are often susceptible to overfitting if they are not meticulously developed and assessed. The existing literature frequently uses a *k*-fold validation to overcome (or minimise) the overfitting issue. Understandably, using training and testing with different proportions of the dataset is likely to reduce overfitting, but it may not be prevented entirely because overfitting is also linked to several other factors, such as improper hyperparameters, sample size, noise in data, and model complexity. Given the multitude of factors associated, overfitting must be thoroughly assessed for a given model besides a *k*-fold validation. To overcome this problem, future research can investigate and monitor the difference in training loss versus validation loss, with the increasing difference between these two as an indicator of overfitting. Another possibility is to apply some overfitting controlling techniques, such as regularisation, early stopping, and a drop-out layer, which have shown promising results in other fields.

5.2.4. *Computational efficiency*

Our review suggests that over 94% of studies did not consider/analyse the computational efficiency of their models, which could be explained as machine learning models used for crash prediction modelling are offline models and they do not need to run in real-time, thereby computational efficiency becomes out of focus (or less important). However, this review emphasises that computational efficiency should be assessed for two reasons. First, real-time crash occurrence prediction models are developed to provide predictions for risky situations in real-time and forecasting in future, so their application appears to be regular for which computational efficiency becomes paramount. Secondly, running complex machine learning-based crash prediction models for a long time is linked to energy loss and environmental degradation (Budennyy et al., 2023). A link between better understanding crash mechanisms and the negative consequences of running these models for a long time is missing. As such, there is a clear need to keep track of energy consumption and equivalent CO_2 emissions for running machine learning-based crash models, which will assist in making decisions between simple and complex models. Also, this monitoring will assist in answering a challenging question: whether the resultant performance gain from a complex model is preferred over a relatively simple model that may have (significantly) lower performance?. Further, with the use of big data for crash modelling, traditional machine learning methods are likely to take a long time to train and provide reasonable performance accuracy, which has a significant impact on the environment. Utilising quantum machine learning methods (e.g., quantum convolutional neural networks) can overcome big data and specialised high-performance computing challenges, which are directly linked to computational efficiency.

5.2.5. *Interpretability of models*

Machine learning models have faced severe criticism for their poor interpretability. With the advent of explainable artificial intelligence, machine learning models have unleashed their potential to provide valuable insights into the causal relationship of a crash with its determinants (Molnar, 2020). Several model-based and post-hoc machine learning interpretation methods have been used and compared (see Section 4 for a summary). Relatively better performance of SHapley Additive exPlanations (SHAP) has been reported in terms of visualising the detailed relationships between a crash and its determinants (Wen et al., 2022).

In essence, these methods allow us to investigate feature ranking/importance, the nature of the relationship (direct or indirect) between an explanatory variable and outcome, dependence, interaction, and clustering. With such information available, some questions for future research are as follows (note that these questions are not linked to any specific interpretability method). Does feature ranking vary significantly with changes in the underlying machine learning model? Can feature ranking and importance obtained from a machine learning model be consistent with a statistical model? How can unobserved heterogeneity be quantified using the methods mentioned above?. Answering these questions will uncover new pathways in further exploring the potential of interpretability techniques in providing meaning to the model output and also assist in bridging the gap between machine learning and statistical models' capabilities.

5.2.6. *Temporal stability of models*

Machine learning-based crash prediction models are developed using the police-reported crash database based on multiple years of data. It has been reported that crash (or injury) determinants exhibit unstable trends over time (Mannering, 2018), which significantly impacts the relationship of a crash with its determinants. Recent statistical and econometric models have addressed this issue by developing separate models corresponding to analysis years and testing the presence of instability using a likelihood ratio test (Mannering, 2018). For machine learning models, separate models can be developed for each analysis year and compare their predictive performance with a model for all analysis years, with the hypothesis that temporal instability will deteriorate model performance. Further, interpretability techniques can also be used to understand whether the impact of any crash factor has changed over time.

5.2.7. *Transferability analysis*

One of the desirable attributes of a machine learning model is to yield similar model performance when tested with different datasets, thereby facilitating generalisation capabilities. Despite such importance of transferability, the transferability of machine learning models has not been well investigated. Transferability analysis should be performed to assess the impact of temporal and spatial variations on model performance. Future research should investigate how a model performs to newer datasets provided that model hyperparameters were tuned to one (or original) dataset. Further, re-tuning hyperparameters with a new dataset should also be performed, and how model performance varies compared to using the same hyperparameters should be assessed. Meanwhile, transfer learning has gained significant attention in recent literature with proven capabilities to yield superior model performance. For instance, Man et al., (2022a) combined Generative Adversarial Network (GAN) and transfer learning to investigate the transferability of real-time crash prediction models. Their study compared the model with two alternatives: standalone models and direct transfer from the baseline model. The findings of this study demonstrate that machine learning models can become transferable temporally, spatially, and spatiotemporally if transfer learning is applied. Whilst preliminary evidence of the efficacy of transfer learning is provided by Man et al., (2022a), rigorous testing with different datasets and for different crash aspects (injury severity and frequency prediction) should be performed to further demonstrate the robustness of transfer learning. Further, the concepts of zero-shot and few-shot learning can be applied to improve the generalisation capabilities of machine learning models to unseen datasets, thereby improving transferability performance. For instance, a transferable graph generation approach based on zero-shot and few-shot learning can be applied, which has been shown to produce promising results in an earlier study (Zhang et al., 2019).

5.2.8. Sensitivity analysis

Crash prediction models aim to provide insights into how the likelihood of a crash or injury varies when explanatory variables are changed. To this end, sensitivity analysis (or marginal effect calculation) is performed to better understand the relationship of outcome with its determinants. One of the methods used for this purpose is local sensitivity analysis, which identifies the marginal effects of an explanatory variable on the outcome, keeping all other variables fixed (at mean or median) (Delen et al., 2006). Several issues persist in local sensitivity analysis, such as strict assumptions of linearity, normality, local variations, and incapability to provide interaction effects (Wen et al., 2021, Wagner, 1995). Given the high non-linearity of machine learning models per se, the insights from local sensitivity analysis may be limited and potentially lead to misleading conclusions (Saltelli and Annoni, 2010).

To overcome these issues, global sensitivity analysis has recently been used in the literature (Jiang et al., 2022), which provides probability distributions as the output of change in explanatory variables that can be linked to one or a set of explanatory variables. Apart from global sensitivity analysis, several methods exist for sensitivity analysis, such as NeuralSens (Pizarroso et al., 2020) and SHAP (Lundberg and Lee, 2017). With these methods available, it becomes necessary to determine whether the results of one method vary from another and, if so, to what extent. Given the unavailability of the ground truth of how the relationship varies between crashes and its determinants, theoretical judgement and empirical evidence complemented by statistical analysis can be utilised to decide which method should be preferred in future research.

5.2.9. Benchmarking framework

With more than two decades of research on machine learning-based crash prediction modelling, the model development and evaluation process appear to be arbitrary rather than systematic, which induces significant bias. Consequently, it remains by and large unclear whether the poor performance of a model is mainly caused by such an arbitrary model development process or the model per se. Accordingly, there is no consensus on assessing these models' performances, with many performance metrics used in past studies. Intuitively, this limitation seriously hinders both the application and improvement of existing models and the development of future models. As such, there is a dire need to develop a benchmarking framework to guide researchers on model development and evaluation process. The underpinning of this framework should be theoretical knowledge about road safety combined with principles of machine learning to develop a sound and robust model. This framework will facilitate objectively comparing the model's strengths and limitations and assist in developing ideas for further improvement in machine learning models.

5.3. Model application

As several machine learning models are developed, it becomes important to understand how these models can be better leveraged for crash prediction modelling and where more research efforts are required, which are highlighted in the ensuing subsections.

5.3.1. Real-time capabilities

An important aspect of the crash occurrence model is its performance outreach for real-time safety assessment. With these models facilitating proactive safety management, one of the goals is to predict crash occurrence in future, which can be used to devise countermeasures to avoid those future crashes. Our literature review suggests that existing models can predict crash occurrence within the next 4–10 mins (Li and Abdel-Aty, 2022a, Zhu et al., 2022, Yuan et al., 2019, Hossain and Muromachi, 2012). However, prediction in such a short window may be less valuable because road management authorities are interested in determining crash occurrence with a longer time buffer (e.g., 30 mins, 60 mins, etc.), so they have enough time to devise mitigation strategies. With machine learning-based powerful forecasting models like recurrent neural network and their variants (e.g., RiskOracle presented by Zhou et al. (2020) and multi-view graph Convolutional Networks proposed by Trirat et al. (2023)), future studies should aim to enhance the future prediction outreach of real-time crash prediction models, which can be used for improving safety in real-time. For instance, a real-time safety model providing pedestrian crash risk forecasting at a signalised intersection can be aided together with signal phasing and timing updates, which can automatically consider crash risk and extend the green time for pedestrians in upcoming signal cycles.

5.3.2. Separate models for different crash types

It is well-known that determinants of one crash type vary significantly compared to those of another crash type. However, very few machine learning models are developed specifically for different crash types, e.g., rear-end, angle (pedestrian crashes), sideswipes, roadway run-off, head-on, among others. The existing literature appears to model many crash types together, hindering understanding of how different factors affect one crash type. Future research should develop separate models for different crash types occurring at different locations because driving behaviour tends to vary significantly in different locations (Ali et al., 2021), which should be considered whilst developing these models.

5.3.3. Bivariate or multivariate modelling

Assessing the safety of a road facility is difficult due to many factors involved, such as different crash types. However, many machine learning models are developed separately for crash frequency, injury severity, or occurrence, with a few exceptions for combining two of them. The underlying reasons could be the volatile nature of crashes, one-way dependency of injury severity on crash occurrence, and differences in response variables (Das and Abdel-Aty, 2011). However, in the authors' view, combining crash frequency and resulting injuries will likely provide a better understanding of safety at a given location. Such combined models will provide insights into the frequency of a given crash and the severity it is likely to produce. To this end, in machine learning models, such as neural networks, the output layer can be hierarchically defined as a combination of regression and classification. Similarly, crash occurrence prediction models can also be linked with injury severity models.

5.3.4. Joint statistical and machine learning models

In a recent review paper, Mannering et al. (2020) reported that "*there is a clear need in the safety field to ground intrinsically predictive models within causal frameworks, while also taking insights from intrinsically predictive models (especially from big data) to improve upon causal structures through insights from associations involving variables not typically available* *in traditional safety data. One promising direction for future research would be a hybrid modeling approach of data-driven and statistical methods (with strong consideration to causal elements)*". Along this line, joint models can be developed, leveraging the benefits of statistical models in explanation and machine learning models in predictions. Further, alternative ways should be devised to better use these two modelling worlds together. For instance, decision trees can be used to provide higher-order interactions for statistical models, and similarly, statistical models can be used to obtain significant explanatory variables for a machine learning model. Nevertheless, the current state-of-the-art on this promising direction appears to be nascent, requiring researchers' significant and urgent attention to fully utilise the potential of joint models.

5.3.5. *Models for connected and automated vehicles*

Our review suggests that models specifically developed for Connected and Automated Vehicles (CAVs) are relatively limited, which can be partially attributed to data scarcity. However, in the last couple of years, data from CAVs have been increasingly available (e.g., Waymo, Argoverse, Lyft, etc.), which can be used to assess their safety. Further, CAVs provide unprecedented opportunities to utilise their massive data to understand safety at a large scale, which is generally hindered by data unavailability. Taking the example of Argoverse, data have been collected for more than 250,000 scenarios of 11 s with an average point cloud of ∼ 107,000 points at a 10 Hz frequency. With such big data, network-wide safety can be assessed, hotspots for a given network can be identified, and their effects on subsequent locations can also be analysed. As such, future research should take advantage of these publicly available databases to develop models tailored for CAVs.

6. Summary and conclusions

Modelling different aspects of a crash, such as crash occurrence, injury severity, and frequency, is critical to understand trends, patterns, devising countermeasures, and assessing their efficacy. To this end, machine learning models have been extensively applied to predict different crash aspects. With more than two decades of research into developing these models and recognising the advancement in big data, the present study aimed at systematically reviewing machine learning models from their first application for crash prediction modelling following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

This paper classified all machine learning models into three broad groups: (a) models for predicting crash occurrence (or real-time crash prediction models), (b) models for predicting crash frequency, and (c) models for predicting injury severity. Then, several modelling intricacies were reviewed, including hyperparameter selection, data sources, sample size, use of explanatory variable selection, and interpretability.

Based on a thorough review, this study pointed out critical gaps in the literature that can guide future research on machine learning model development. Critical research needs are categorised into three aspects: model development, evaluation and application. Although significant efforts have been dedicated to model development, exigent issues like capturing unobserved heterogeneity, spatiotemporal correlation, computational efficiency, and class imbalance have received rather limited attention. Whilst models for some frequent collision types have been developed, some critical interactions, such as vehicle-bicycle and automated vehicles-pedestrian, have been largely ignored. Finally, our review suggested that machine learning-based crash frequency models are predominantly tree-based, with some exceptions where neural networks are used. More research is required to further leverage the power of neural networks, especially where they appear to be most useful, e.g., forecasting or trend predictions.

Recognising the complexity of crash modelling, the application of advanced machine learning methods appears to be at a nascent stage. To this end, this review recommended using advanced algorithms and architectures capable of handling complexity and providing high performance accuracy. For instance, for real-time or near-real-time predictions advanced architectures like temporal convolutional networks, probabilistic neural networks, or attention-based models could be used. These architectures are designed to capture long-term dependencies in time-series data, which is crucial for accurate real-time predictions. The use of hybrid deep learning models for real-time crash occurrence is recommended, which could include combination of different machine learning models to improve model performance. For instance, a deep spatiotemporal hybrid network can be used that integrates a Convolutional Neural Network, Long Short-term Memory, and Artificial Neural Network to incorporate the synergistic power of individual models for real-time crash occurrence (Kashifi et al., 2023). Another possible application of deep hybrid network could be grid-based real-time crash prediction (see some preliminary evidence in Kashifi et al. (2022)). Whilst the use of self-supervised learning appears to be minimal in real-time crash occurrence context, it can be used to accurately annotate non-crash events, which can learn semantically meaningful feature representations via pre-crash driving conditions that do not require semantic annotations. Recognising the presence of time series nature in crash and non-crash data, neural ordinary differential equations can also be applied, which are found to be powerful in modelling time series data in different fields like finance, healthcare, and environmental monitoring.

Similarly, for crash frequency predictions, Generative Adversarial Networks or Variational Autoencoders could offer more nuanced insights by modelling the underlying data distribution more effectively. These architectures are particularly useful for capturing spatio-temporal patterns, which are often critical in predicting crash frequencies. Along this line, capsule networks can be applied, which are specifically designed to capture hierarchical and spatial relationships in crash frequency models as noted by several studies in the literature (Zeng et al., 2020, Wang and Huang, 2016, Ahmed et al., 2011).

Finally, the use of specialised classification techniques for predicting injury severity is warranted, which could include (i) ordinal regression neural networks that models the ordinal nature of injury severity categories more effectively, and (ii) transformer-based models that could handle imbalanced datasets and high-dimensional feature spaces. Along this direction, a bidirectional encoder representations-based transformers model (Oliaee et al., 2023) can be developed to classify traffic injury severities using crash narrative reports — a common way of recording injury severity data. Further, graph-based models can also be used to represent the relationships between injury severity classes as a graph, whereby graph convolutional networks can be applied to capture information and make injury severity prediction. Since, unobserved heterogeneity is inevitable in crash data, the use of multimodal data (e.g., crash records with trajectories obtained from sensors, video imagery, and text) and advanced methods like Neural Graphical Models and Deep Belief Network is recommended. Meanwhile, for modelling a network level, graph neural networks are suitable that join links and nodes, which could be useful in crash modelling of midblocks and intersections simultaneously or corridor-wide safety.

Diffusion models are another alternative that is used for generative tasks and modelling data distributions. Whilst their direct application in crash modelling is still vague, it can assist during different machine learning model development phases, e.g., data pre-processing and denoising crash data, feature generation especially in scarce crash data, and data augmentation for generalisation of crash prediction models.

Although this paper reviews machine learning methods for predicting different crash aspects, it is worth noting that using machine learning techniques may not be necessary in all cases. For instance, to understand causal relationships and inferences, machine learning methods are still evolving, and therefore alternative and more mature casual inference methodologies (e.g., casual Markov chains) should be used.

It is envisioned that this review will inspire methodological advancements in machine learning-based crash prediction models, which

Table A1
Summary of representative machine learning models used for crash prediction modelling.

Model class	Model	Characteristics	Challenges
Decision tree-based	Classification and regression tree	• Simple, computationally efficient, and can handle non-linearity Uses rules to arrive at a decision Can be used to determine the contribution of explanatory variables Implicit feature selection Less sensitive to outliers	• Susceptible to data/class imbalance Binary split restriction Prone to overfitting High variance and low bias With a growing number of classes, the tree may become unstable
	Random Forest	• Robust to overfitting Less sensitive to outliers Inherent ability to rank the importance of explanatory variables	• Difficult to interpret Involves trade-off between training time and increased number of trees Poor performance for small data
	Adaptive Boosting	• Combines weak learners to produce a strong classifier Less prone to overfitting Can handle missing values and outliers	• Cannot extract linear combination of explanatory variables High variance Slower in training
	Gradient Boosted Decision Tree	• Flexible in optimising loss function and tuning hyperparameters Can handle missing values	• Minimising errors may lead to overemphasising the outliers Prone to overfitting Less computationally efficient Less interpretative
	Extreme Gradient Boosting Tree	• Computationally efficient Can handle missing values	• Sensitive to outliers Difficult to scale up
Neural network-based	Artificial Neural Network	• Ability to learn and model non-linear and complex relationships No restrictions on the input variables Caters for data with high volatility and non-constant variance	• Prone to overfitting Less generalisation capability Lack of interpretability Fall into local minima
	Convolutional Neural Network	• Inherent capability to identify important explanatory variables Weight sharing Best for high-dimensional (or image) data	• Huge training data is required Time-consuming to train
	Recurrent Neural Network (RNN)/Long Short-Term Memory (LSTM)	RNN Capture short-term dependency among time series data Weight sharing across time steps LSTM Can handle long-term dependencies Less sensitive to vanishing gradient problem Complex sequential data	RNN Vanishing gradient issue Exploding gradient issue Time-consuming to train Some activation functions cannot process long sequence LSTM Complicated than RNNs Require more training data Time-consuming to train for large data
Others	Support Vector Machine	• Computationally efficient Can find global minima Mitigate overfitting Best for binary classification	• Cannot inherently select/rank explanatory variables Sensitive to hyperparameters (kernel functions and parameters) Difficult to interpret
	Naïve Bayes	• Can handle both binary and multi-level classification A small amount of training data is required Fast and efficient	• Sensitive to numerical variables Strong feature independence assumption Zero frequency issue
	Bayesian Networks	• Graphic representations of complex structures No prior assumptions Identifying the causal relationship between variables	• Time-consuming to train Plagued with heuristics and biases Formal logic is not necessarily a good model of human reasoning
	K-Nearest Neighbour	• Simple to implement Can handle missing values Capture non-linearity	• Computationally less efficient High memory storage is required Sensitive to irrelevant features

is the intersection of safety researchers and computer scientists. Future research should focus on developing models with high predictive power and interpretability, which requires methodologies for combining statistical and machine learning models.

CRediT authorship contribution statement

Yasir Ali: Conceptualization, Methodology, Software, Data curation, Formal analysis, Writing – original draft, Visualization. **Fizza Hussain:** Conceptualization, Writing – review & editing. **Md Mazharul Haque:** Methodology, Investigation, Writing – review & editing.

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

Data will be made available on request.

Acknowledgement

The authors are grateful to the four reviewers whose comments have greatly improved the manuscript. We also would like to thank

researchers from computer/data science who shared their insights into advanced modelling methodologies for crash prediction.

Appendix

Appendix A. Supplementary data

A database file can be found online containing the full list of studies along with details like outcome of interest (frequency, severity etc.), train-test ratio, data, and machine learning method applied. Supplementary data to this article can be found online at https://doi.org/10.1016/j.aap.2023.107378.

References

Abay, K.A., 2015. Investigating the nature and impact of reporting bias in road crash data. Transp. Res. A Policy Pract. 71, 31–45.

Abdelwahab, H.T., Abdel-Aty, M.A., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transp. Res. Rec. 1746, 6–13.

Abou Elassad, Z.E., Mousannif, H., Al Moatassime, H., 2020a. A proactive decision support system for predicting traffic crash events: a critical analysis of imbalanced class distribution. Knowl.-Based Syst. 205, 106314.

Abou Elassad, Z.E., Mousannif, H., Al Moatassime, H., 2020b. A real-time crash prediction fusion framework: an imbalance-aware strategy for collision avoidance systems. Transp. Res. Part C: Emerg. Tech. 118, 102708.

Ahmad, N., Wali, B., Khattak, A.J., 2022. Heterogeneous ensemble learning for enhanced crash forecasts–a frequentist and machine learning based stacking framework. J. Saf. Res.

Ahmed, M., Huang, H., Abdel-Aty, M., Guevara, B., 2011. Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. Accid. Anal. Prev. 43, 1581–1589.

Ali, Y., Haque, M.M., Zheng, Z., Bliemer, M.C., 2021. Stop or go decisions at the onset of yellow light in a connected environment: A hybrid approach of decision tree and panel mixed logit model. Anal. Methods Accident Res. 31, 100165.

Ali, Y., Bliemer, M.C., Haque, M.M., Zheng, Z., 2022. Examining braking behaviour during failed lane-changing attempts in a simulated connected environment with driving aids. Transp. Res. Part C: Emerg. Tech. 136, 103531.

Ali, Y., Haque, M.M., Mannering, F., 2023a. Assessing traffic conflict/crash relationships with extreme value theory: recent developments and future directions for connected and autonomous vehicle and highway safety research. Anal. Methods Accident Res. 100276.

Ali, Y., Haque, M.M., Mannering, F., 2023b. A Bayesian generalised extreme value model to estimate real-time pedestrian crash risks at signalised intersections using artificial intelligence-based video analytics. Anal. Methods Accident Res. 38, 100264.

Almamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R. and Frefer, A. A. Comparison of machine learning algorithms for predicting traffic accident severity. 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT), 2019. IEEE, 272-276.

Amoros, E., Martin, J.-L., Laumon, B., 2006. Under-reporting of road crash casualties in France. Accid. Anal. Prev. 38, 627–635.

Angarita-Zapata, J.S., Maestre-Gongora, G., Calderín, J.F., 2021. A bibliometric analysis and benchmark of machine learning and automl in crash severity prediction: the case study of three colombian cities. Sensors 21, 8401.

Arteaga, C., Paz, A., Park, J., 2020. Injury severity on traffic crashes: a text mining with an interpretable machine-learning approach. Saf. Sci. 132, 104988.

Assi, K., 2020. Traffic crash severity prediction—A synergy by hybrid principal component analysis and machine learning models. Int. J. Environ. Res. Public Health 17, 7598.

Bao, J., Liu, P., Ukkusuri, S.V., 2019. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. Accid. Anal. Prev. 122, 239–254.

Basso, F., Basso, L.J., Bravo, F., Pezoa, R., 2018. Real-time crash prediction in an urban expressway using disaggregated data. Transp. Res. Part C: Emerg. Tech. 86, 202–219.

Basso, F., Basso, L.J., Pezoa, R., 2020. The importance of flow composition in real-time crash prediction. Accid. Anal. Prev. 137, 105436.

Basso, F., Pezoa, R., Varas, M., Villalobos, M., 2021. A deep learning approach for real-time crash prediction using vehicle-by-vehicle data. Accid. Anal. Prev. 162, 106409.

Blincoe, L.J., Seay, A.G., Zaloshnja, E., Miller, T.R., Romano, E.O., Luchter, S., Spicer, R. S., 2002. The economic impact of motor vehicle crashes, 2000. United States, National Highway Traffic Safety Administration.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Browne, M.W., 2000. Cross-validation methods. J. Math. Psychol. 44, 108–132.

Budennyy, S., Lazarev, V., Zakharenko, N., Korovin, A., Plosskaya, O., Dimitrov, D., Akhripkin, V., Pavlov, I., Oseledets, I. and Barsola, I. Eco2ai: carbon emissions tracking of machine learning models as the first step towards sustainable ai. Doklady Mathematics, 2023. Springer, 1-11.

Cai, Q., Abdel-Aty, M., Yuan, J., Lee, J., Wu, Y., 2020. Real-time crash prediction on expressways using deep generative models. Transp. Res. Part C: Emerg. Tech. 117, 102697.

Candefjord, S., Muhammad, A.S., Bangalore, P., Buendia, R., 2021. On scene injury severity prediction (OSISP) machine learning algorithms for motor vehicle crash occupants in US. J. Transp. Health 22, 101124.

Chang, L.-Y., Chen, W.-C., 2005. Data mining of tree-based models to analyze freeway accident frequency. J. Saf. Res. 36, 365–375.

Chang, L.-Y., Wang, H.-W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid. Anal. Prev. 38, 1019–1027.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016. 785-794.

Chen, C., Zhang, G., Tarefder, R., Ma, J., Wei, H., Guan, H., 2015. A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. Accid. Anal. Prev. 80, 76–88.

Chen, C., Zhang, G., Qian, Z., Tarefder, R.A., Tian, Z., 2016a. Investigating driver injury severity patterns in rollover crashes using support vector machine models. Accid. Anal. Prev. 90, 128–139.

Chen, C., Zhang, G., Yang, J., Milton, J.C., 2016b. An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. Accid. Anal. Prev. 90, 95–107.

Das, A., Abdel-Aty, M.A., 2011. A combined frequency–severity approach for the analysis of rear-end crashes on urban arterials. Saf. Sci. 49, 1156–1163.

Das, S., Le, M., Dai, B., 2020. Application of machine learning tools in classifying pedestrian crash types: a case study. Transp. Safety Environ. 2, 106–119.

De Oña, J., Mujalli, R.O., Calvo, F.J., 2011. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Accid. Anal. Prev. 43, 402–411.

De Oña, J., López, G., Abellán, J., 2013. Extracting decision rules from police accident reports through decision trees. Accid. Anal. Prev. 50, 1151–1160.

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accid. Anal. Prev. 38, 434–444.

DfT. 2011. Collision Recording and SHaring (CRASH) [Online]. Available: http://webarchive.nationalarchives.gov.uk/20110503151558/http:/dft.gov.uk/pgr/statistics/committeesusergroups/crash [Accessed 26 January 2023].

Ding, C., Chen, P., Jiao, J., 2018. Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: a machine learning approach. Accid. Anal. Prev. 112, 116–126.

Dong, S., Khattak, A., Ullah, I., Zhou, J., Hussain, A., 2022. Predicting and analyzing road traffic injury severity using boosting-based ensemble learning models with SHAPley Additive exPlanations. Int. J. Environ. Res. Public Health 19, 2925.

Fang, J., Qiao, J., Bai, J., Yu, H., Xue, J., 2022. Traffic accident detection via self-supervised consistency learning in driving scenarios. IEEE Trans. Intell. Transp. Syst. 23, 9601–9614.

Ghandour, A.J., Hammoud, H., Al-Hajj, S., 2020. Analyzing factors associated with fatal road crashes: a machine learning approach. Int. J. Environ. Res. Public Health 17, 4111.

Goswamy, A., Abdel-Aty, M., Islam, Z., 2023. Factors affecting injury severity at pedestrian crossing locations with Rectangular RAPID Flashing Beacons (RRFB) using XGBoost and random parameters discrete outcome models. Accid. Anal. Prev. 181, 106937.

Haque, M.M., Ohlhauser, A.D., Washington, S., Boyle, L.N., 2016. Decisions and actions of distracted drivers at the onset of yellow lights. Accid. Anal. Prev. 96, 290–299.

He, S., Sadeghi, M. A., Chawla, S., Alizadeh, M., Balakrishnan, H. and Madden, S. Inferring high-resolution traffic accident risk maps based on satellite imagery and GPS trajectories. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 11977-11985.

Hossain, M., Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accid. Anal. Prev. 45, 373–381.

Hu, J., Huang, M.-C., Yu, X., 2020. Efficient mapping of crash risk at intersections with connected vehicle data and deep learning models. Accid. Anal. Prev. 144, 105665.

Hu, Z., Shi, Q., Chen, Y., Yuan, Q., Tao, Z., Bian, Y., Haque, M.M., 2022. Analyzing factors and interaction terms affecting urban fatal crash types based on a hybrid framework of econometric model and machine learning approaches. Int. J. Crashworthiness 1–13.

Hussain, F., Li, Y., Arun, A., Haque, M.M., 2022. A hybrid modelling framework of machine learning and extreme value theory for crash risk estimation using traffic conflicts. Anal. Methods Accident Res. 36, 100248.

Ihueze, C.C., Onwurah, U.O., 2018. Road traffic accidents prediction modelling: an analysis of Anambra State, Nigeria. Accid. Anal. Prev. 112, 21–29.

Ijaz, M., Zahid, M., Jamal, A., 2021. A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw. Accid. Anal. Prev. 154, 106094.

Imprialou, M., Quddus, M., 2019. Crash data quality for road safety research: current state and future directions. Accid. Anal. Prev. 130, 84–90.

Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accid. Anal. Prev. 108, 27–36.

Islam, Z., Abdel-Aty, M., Cai, Q., Yuan, J., 2021. Crash data augmentation using variational autoencoder. Accid. Anal. Prev. 151, 105950.

Janstrup, K.H., Kaplan, S., Hels, T., Lauritsen, J., Prato, C.G., 2016. Understanding traffic crash under-reporting: linking police and medical records to individual and crash characteristics. Traffic Inj. Prev. 17, 580–584.

Jeong, H., Jang, Y., Bowman, P.J., Masoud, N., 2018. Classification of motor vehicle crash injury severity: a hybrid approach for imbalanced data. Accid. Anal. Prev. 120, 250–261.

Ji, A., Levinson, D., 2020. Injury severity prediction from two-vehicle crash mechanisms with machine learning and ensemble models. IEEE Open J. Intell. Transp. Syst. 1, 217–226.

Jiang, L., Xie, Y., Wen, X., Ren, T., 2022. Modeling highly imbalanced crash severity data by ensemble methods and global sensitivity analysis. J. Transp. Safety Security 14, 562–584.

Jiang, F., Yuen, K.K.R., Lee, E.W.M., 2020. A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions. Accid. Anal. Prev. 141, 105520.

Karim, M.M., Li, Y., Qin, R., Yin, Z., 2022. A dynamic spatial-temporal attention network for early anticipation of traffic accidents. IEEE Trans. Intell. Transp. Syst. 23, 9590–9600.

Kashifi, M.T., Al-Sghan, I.Y., Rahman, S.M., Al-Ahmadi, H.M., 2022. Spatiotemporal grid-based crash prediction—application of a transparent deep hybrid modeling framework. Neural Comput. & Applic. 34, 20655–20669.

Kashifi, M.T., Al-Turki, M., Sharify, A.W., 2023. Deep hybrid learning framework for spatiotemporal crash prediction using big traffic data. Int. J. Transp. Sci. Technol. 12, 793–808.

Li, P., Abdel-Aty, M., 2022a. A hybrid machine learning model for predicting real-time secondary crash likelihood. Accid. Anal. Prev. 165, 106504.

Li, P., Abdel-Aty, M., 2022b. Real-time crash likelihood prediction using temporal attention-based deep learning and trajectory fusion. J. Transp. Eng., Part A: Systems 148, 04022043.

Li, Z., Liu, P., Wang, W., Xu, C., 2012. Using support vector machine models for crash injury severity analysis. Accid. Anal. Prev. 45, 478–486.

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accid. Anal. Prev. 40, 1611–1618.

Li, L., Sheng, X., Du, B., Wang, Y., Ran, B., 2020. A deep fusion model based on restricted Boltzmann machines for traffic accident duration prediction. Eng. Appl. Artif. Intel. 93, 103686.

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann. Intern. Med. 151, W-65-W-94.

Lin, D.-J., Chen, M.-Y., Chiang, H.-S., Sharma, P.K., 2021. Intelligent traffic accident prediction model for Internet of Vehicles with deep learning approach. IEEE Trans. Intell. Transp. Syst. 23, 2340–2349.

Lin, M.-R., Kraus, J.F., 2008. Methodological issues in motorcycle injury epidemiology. Accid. Anal. Prev. 40, 1653–1660.

Lu, P., Zheng, Z., Ren, Y., Zhou, X., Keramati, A., Tolliver, D. and Huang, Y. 2020. A gradient boosting crash prediction approach for highway-rail grade crossing crash analysis. Journal of advanced transportation, 2020.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Proces. Syst. 30.

Ma, X., Lu, J., Liu, X., Qu, W., 2022. A genetic programming approach for real-time crash prediction to solve trade-off between interpretability and accuracy. J. Transp. Safety Security 1–23.

Man, C.K., Quddus, M., Theofilatos, A., 2022a. Transfer learning for spatio-temporal transferability of real-time crash prediction models. Accid. Anal. Prev. 165, 106511.

Man, C.K., Quddus, M., Theofilatos, A., Yu, R., Imprialou, M., 2022b. Wasserstein generative adversarial network to address the imbalanced data problem in real-time crash risk prediction. IEEE Trans. Intell. Transp. Syst. 23, 23002–23013.

Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Anal. Methods Accident Res. 17, 1–13.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Anal. Methods Accident Res. 1, 1–22.

Mannering, F., Bhat, C.R., Shankar, V., Abdel-Aty, M., 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Anal. Methods Accident Res. 25, 100113.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accident Res. 11, 1–16.

Mease, D., Wyner, A.J., Buja, A., 2007. Boosted classification trees and class probability/quantile estimation. J. Mach. Learn. Res. 8.

Methley, A.M., Campbell, S., Chew-Graham, C., Mcnally, R., Cheraghi-Sohi, S., 2014. PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. BMC Health Serv. Res. 14, 1–10.

Mokhtarimousavi, S., Anderson, J.C., Azizinamini, A., Hadi, M., 2019. Improved support vector machine models for work zone crash injury severity prediction and analysis. Transp. Res. Rec. 2673, 680–692.

Mokhtarimousavi, S., Anderson, J.C., Hadi, M., Azizinamini, A., 2021. A temporal investigation of crash severity factors in worker-involved work zone crashes: Random parameters and machine learning approaches. Transp. Res. Interdisciplinary Perspectives 10, 100378.

Molnar, C., 2020. Interpretable machine learning. Lulu. com.

Montella, A., Chiaradonna, S., Criscuolo, G., De Martino, S., 2019. Development and evaluation of a web-based software for crash data collection, processing and analysis. Accid. Anal. Prev. 130, 108–116.

Morris, C., Yang, J.J., 2021. Effectiveness of resampling methods in coping with imbalanced crash data: crash type analysis and predictive modeling. Accid. Anal. Prev. 159, 106240.

Nikolaou, D., Ziakopoulos, A., Dragomanovits, A., Roussou, J., Yannis, G., 2023. Comparing machine learning techniques for predictions of motorway segment crash risk level. Safety 9, 32.

Ogle, J.H., 2007. Technologies for improving safety data. Transportation Research Board.

Oliaee, A.H., Das, S., Liu, J., Rahman, M.A., 2023. Using bidirectional encoder representations from transformers (BERT) to classify traffic crash severity types. Natural Language Processing Journal 3, 100007.

Parsa, A.B., Taghipour, H., Derrible, S., Mohammadian, A.K., 2019. Real-time accident detection: coping with imbalanced data. Accid. Anal. Prev. 129, 202–210.

Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S., Mohammadian, A.K., 2020. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accid. Anal. Prev. 136, 105405.

Picard, R.R., Cook, R.D., 1984. Cross-validation of regression models. J. Am. Stat. Assoc. 79, 575–583.

Pizarroso, J., Portela, J. and Muñoz, A. 2020. NeuralSens: sensitivity analysis of neural networks. arXiv preprint arXiv:2002.11423.

Quddus, M.A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. Accid. Anal. Prev. 40, 1486–1497.

Ramotowski, M., Fitzgerald, R., 2020. Chi-Squared Automatic Inference Detection (CHAID) decision tree. Apache Software License, Version, p. 5.

Ribeiro, M. T., Singh, S. and Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016. 1135-1144.

Saltelli, A., Annoni, P., 2010. How to avoid a perfunctory sensitivity analysis. Environ. Model. Softw. 25, 1508–1517.

Saltelli, A., Tarantola, S., Chan, K.-S., 1999. A quantitative model-independent method for global sensitivity analysis of model output. Technometrics 41, 39–56.

Santos, K., Dias, J.P., Amado, C., 2022. A literature review of machine learning algorithms for crash injury severity prediction. J. Saf. Res. 80, 254–269.

Sarkis-Onofre, R., Catalá-López, F., Aromataris, E., Lockwood, C., 2021. How to properly use the PRISMA Statement. Syst. Rev. 10, 1–3.

Sattar, K., Chikh Oughali, F., Assi, K., Ratrout, N., Jamal, A., Masiur Rahman, S., 2023. Transparent deep machine learning framework for predicting traffic crash severity. Neural Comput. & Applic. 35, 1535–1547.

Schlögl, M., Stütz, R., Laaha, G., Melcher, M., 2019. A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset. Accid. Anal. Prev. 127, 134–149.

Shrivastava, H. and Chajewska, U. 2022. Neural graphical models. arXiv preprint arXiv:2210.00453.

Silva, P.B., Andrade, M., Ferreira, S., 2020. Machine learning applied to road safety modeling: a systematic literature review. J. Traffic Transp. Eng. (english Edition) 7, 775–790.

Sinha, A., Vu, V., Chand, S., Wijayaratna, K., Dixit, V., 2021. A crash injury model involving autonomous vehicle: Investigating of crash and disengagement reports. Sustainability 13, 7938.

Sun, J., Sun, J., Chen, P., 2014. Use of support vector machine models for real-time prediction of crash risk on urban expressways. Transp. Res. Rec. 2432, 91–98.

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.

Theofilatos, A., Chen, C., Antoniou, C., 2019. Comparing machine learning and deep learning methods for real-time crash prediction. Transp. Res. Rec. 2673, 169–178.

Trirat, P., Yoon, S., Lee, J.-G., 2023. MG-TAR: multi-view graph convolutional networks for traffic accident risk prediction. IEEE Trans. Intell. Transp. Syst. 24, 3779–3794.

Vapnik, V., 1999. The nature of statistical learning theory. Springer science & business media.

Wagner, H.M., 1995. Global sensitivity analysis. Oper. Res. 43, 948–969.

Wang, J., Huang, H., 2016. Road network safety evaluation using Bayesian hierarchical joint model. Accid. Anal. Prev. 90, 152–158.

Watson, A., Watson, B., Vallmuur, K., 2015. Estimating under-reporting of road crash injuries to police using multiple linked data collections. Accid. Anal. Prev. 83, 18–25.

Wei, Z., Zhang, Y., Das, S., 2022. Applying explainable machine learning techniques in daily crash occurrence and severity modeling for rural interstates. Transp. Res. Rec. 03611981221134629.

Wen, X., Xie, Y., Jiang, L., Pu, Z., Ge, T., 2021. Applications of machine learning methods in traffic crash severity modelling: current status and future directions. Transp. Rev. 41, 855–879.

Wen, X., Xie, Y., Jiang, L., Li, Y., Ge, T., 2022. On the interpretability of machine learning methods in crash frequency modeling and crash modification factor development. Accid. Anal. Prev. 168, 106617.

WHO. 2023. Road traffic injuries [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries [Accessed 11 January 2023].

Wohlin, C., 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, pp. 1–10.

Yang, L., Aghaabbasi, M., Ali, M., Jan, A., Bouallegue, B., Javed, M.F., Salem, N.M., 2022. Comparative analysis of the optimized KNN, SVM, and ensemble DT models using Bayesian optimization for predicting pedestrian fatalities: an advance towards realizing the sustainable safety of pedestrians. Sustainability 14, 10467.

Yannis, G., Papadimitriou, E., Chaziris, A., Broughton, J., 2014. Modeling road accident injury under-reporting in Europe. Eur. Transp. Res. Rev. 6, 425–438.

Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prev. 51, 252–259.

Yu, R., Abdel-Aty, M., 2014. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Saf. Sci. 63, 50–56.

Yuan, J., Abdel-Aty, M., Gong, Y., Cai, Q., 2019. Real-time crash risk prediction using long short-term memory recurrent neural network. Transp. Res. Rec. 2673, 314–326.

Zarei, M., Hellinga, B., Izadpanah, P., 2023. Application of Conditional Deep Generative Networks (CGAN) in empirical bayes estimation of road crash risk and identifying crash hotspots. Int. J. Transp. Sci. Technol.

Zeng, Q., Wen, H., Huang, H., Wang, J., Lee, J., 2020. Analysis of crash frequency using a Bayesian underreporting count model with spatial correlation. Physica A 545, 123754.

Zhang, S., Abdel-Aty, M., 2022. Real-time crash potential prediction on freeways using connected vehicle data. Analytic Methods in Accident Research 36, 100239.

Zhang, S., Khattak, A., Matara, C.M., Hussain, A., Farooq, A., 2022a. Hybrid feature selection-based machine learning Classification system for the prediction of injury severity in single and multiple-vehicle accidents. PLoS One 17, e0262941.

Zhang, C., Lyu, X., Tang, Z.T.G.G., 2019. Transferable graph generation for zero-shot and few-shot learning. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 1641–1649.

Zhang, Z., Nie, Q., Liu, J., Hainen, A., Islam, N., Yang, C., 2022b. Machine learning based real-time prediction of freeway crash risk using crowdsourced probe vehicle data. J. Intell. Transp. Syst. 1–19.

Zhou, D., Gayah, V.V., Wood, J.S., 2022. Integration of machine learning and statistical models for crash frequency modeling. Transportation Letters 1–12.

Zhou, Z., Wang, Y., Xie, X., Chen, L., Liu, H.R., 2020. A minute-level citywide traffic accident forecasting framework. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1258–1265.

Zhu, M., Yang, H.F., Liu, C., Pu, Z., Wang, Y., 2022. Real-time crash identification using connected electric vehicle operation data. Accid. Anal. Prev. 173, 106708.