# Road Traffic Accident Prediction Based on Multi-Source Data – A Systematic Review

Meiling HE[1], Guangrong MENG [2], Xiaohui WU[3], Xun HAN[4], Jiangyang FAN[5]

Original Scientific Paper
Submitted: 27 May 2024
Accepted: 23 Sep 2024

[1] hemeiling@ujs.edu.cn, Jiangsu University, School of Automotive and Traffic Engineering
[2] 2222204167@stmail.ujs.edu.cn, Jiangsu University, School of Automotive and Traffic Engineering
[3] wuxiaohui@ujs.edu.cn, Jiangsu University, School of Automotive and Traffic Engineering
[4] Corresponding author, hanxun@scpolicec.edu.cn, Sichuan Police College, Intelligent Policing Key Laboratory of Sichuan Province
[5] 3210409049@stmail.ujs.edu.cn, Jiangsu University, School of Automotive and Traffic Engineering

Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

**ABSTRACT**

With the acceleration of urbanisation and the rapid increase in road traffic volume, the scientific prediction of traffic accidents has become crucial for improving road safety and enhancing traffic efficiency. However, traffic accident prediction is a complex and multifaceted problem that requires the comprehensive consideration of multiple factors, including people, vehicles, roads and the environment. This paper provides a detailed analysis of traffic accident prediction based on multi-source data. By thoroughly considering data sources, data processing and prediction methods, this paper introduces the various aspects of traffic accident prediction from different perspectives. It helps readers understand the characteristics of different data and methods, the process of accident prediction and the key technologies involved. At the end of the paper, the main challenges and future directions in road crash prediction research are summarised. For example, the lack of efficient data sharing between different departments and fields poses significant challenges to the integration of multi-source data. In the future, combining deep learning models with time-sensitive data, such as social media and vehicle network data, could effectively improve the accuracy of real-time accident prediction.

**KEYWORDS**

multi-source data; road traffic accident; data processing; statistical learning; machine learning; deep learning.

## 1. INTRODUCTION

### 1.1 Research background

Road traffic injuries represent a major global public health challenge, causing millions of deaths and disabilities annually, along with significant economic and social losses. In September 2020, the United Nations General Assembly adopted Resolution A/RES/74/299, titled “Improving Global Road Safety”, which launched the “Decade of Action for Road Safety 2021–2030”. The goal is to reduce road traffic fatalities and injuries by 50% by 2030 [1]. According to the WHO’s “Global Status Report on Road Safety 2023”, global road traffic fatalities slightly decreased to 1.19 million in 2023 compared to 2022 but remained alarmingly high. Traffic injuries are still the leading cause of death among individuals aged 5–29, particularly in low- and middle-income countries, where pedestrians and cyclists account for over half of the fatalities [2].

Traffic accident prediction technologies play a critical role in reducing road traffic fatalities and injuries. Leveraging advanced data analysis and machine learning algorithms, high-risk areas and periods can be identified, enabling targeted preventive interventions that lower accident rates and casualties, thereby improving road safety.

The development of intelligent transportation systems and advancements in IoT have diversified traffic accident prediction by integrating multi-source data. Traffic cameras, GPS devices, smartphone sensors, weather data, social media and dashcams now provide comprehensive insights into traffic conditions, driving behaviour and environmental factors. Utilising these data sources enhances prediction accuracy and timeliness, facilitating proactive traffic management that reduces accidents and casualties, thereby contributing to a safer road environment.

### 1.2 Research gap

Traffic accident prediction relies on historical data and models to estimate the likelihood of accidents occurring in specific space-time regions. This multi-faceted problem demands a comprehensive analysis of factors such as human behaviour, vehicles, roads and environmental conditions to produce accurate predictions [3-5]. Given that these factors may originate from diverse institutions, facilities or datasets, acquiring data from multiple sources is essential for precise predictions [6]. This article provides a detailed summary and analysis of research on traffic accident prediction using multi-source data, focusing on three main steps: data acquisition, data processing and prediction models.

Traffic accidents typically result from an interaction of factors related to people, vehicles, roads and the environment. Earlier research mainly focused on vehicle dynamics, analysing lateral and longitudinal movements along with horizontal [7, 8] and vertical indicators [9-11] to evaluate road risk. However, environmental factors were often overlooked. Subsequent studies have systematically reviewed how traffic and road characteristics contribute to accidents [12]. To enhance prediction accuracy, it is crucial to integrate multiple data sources, as each source offers unique insights. Researchers advocate for the inclusion of heterogeneous data and consideration of data quality in model accuracy [13]. They also highlight the importance of utilising large-scale datasets, addressing spatial heterogeneity, and managing high-dimensional data to enable real-time prediction. Nonetheless, previous studies have not fully addressed common challenges like missing data, outliers and imbalanced datasets, which this paper aims to rectify.

The selection of an appropriate prediction model, based on different datasets and data processing techniques, is crucial to the accuracy of traffic accident prediction. The development of traffic accident prediction models has gone through three stages: statistical learning, machine learning and deep learning. Statistical learning methods are suitable for analysing correlations and trends between data [14, 15], machine learning excels at uncovering complex patterns and non-linear relationships in the data [16-18], while deep learning can handle large-scale, high-dimensional data and autonomously learn complex nonlinear patterns [19]. In recent years, neural networks have played a significant role in identifying and describing the factors influencing the frequency and severity of road accidents [20]. Existing studies often focus on accident models for specific prediction scenarios but lack a systematic introduction to the characteristics and applicable scenarios of the main models across the three developmental stages of traffic accident prediction methods, which is one of the key distinctions of this paper.

### 1.3 Objectives and contributions

Addressing the shortcomings in previous research on traffic accident prediction, this paper systematically examines traffic accident data sources, data processing methods, predictive models and potential future developments. It provides a comprehensive review and summary of the key aspects involved in traffic accident prediction.

This paper makes the following main contributions:

1) This paper presents a multifaceted overview of traffic accident prediction, aiming to elucidate the advantages of various data sources and methodologies, the prediction process and key technologies involved.
2) The prediction methods are categorised into three types: statistical learning, machine learning and deep learning. For each category, specific models are reviewed, comparing their characteristics, prediction objectives and applicable datasets.
3) The paper also addresses the primary challenges and future research directions in predicting traffic accidents using multi-source data.

### 1.4 Research process

To ensure the high quality and relevance of the research, we retrieved articles from databases such as Web of Science, Engineering Village (EI), Science Direct and IEEE, focusing primarily on English journals and conference papers published after 1 January 2010. We then reviewed the selected papers individually, excluding those that do not contain relevant technical content or lack practical significance.

To conduct literature retrieval and selection, this study employs various search strings with keywords and Boolean operators "AND/OR" for identifying relevant research. The * denotes a wildcard, for example, "Forec*" matches "forecast", "forecasted", "forecasting", etc.

(1) Research data sources: ("traffic accident" OR "traffic crash" OR "traffic collision" OR "crash risk") AND (predict* OR forec* OR detect*) AND "data source" AND (government dat* OR open dat* OR sensor dat* OR social media dat* OR geospatial dat* OR private sector dat* OR crowdsourcing dat*).

(2) Data processing: "data process*" AND ("data cleaning" OR "feature extraction" OR "standardisation" AND "normalisation" OR "data balancing").

(3) Research methods: ("traffic accident" OR "traffic crash" OR "traffic collision" OR "crash risk") AND (predict* OR forec* OR detect*) AND (method* OR algorithm*) AND (statistical learning OR machine learning OR deep learning).

Using the aforementioned keywords, a search was conducted in the Web of Science Core Collection database to identify research on traffic accidents and related hotspots from 1 January 2010 to the present. *Figure 1* demonstrates that research on traffic accident prediction primarily focuses on prediction models and road safety. In recent years, the use of deep learning for accident prediction has emerged as a prominent research topic.

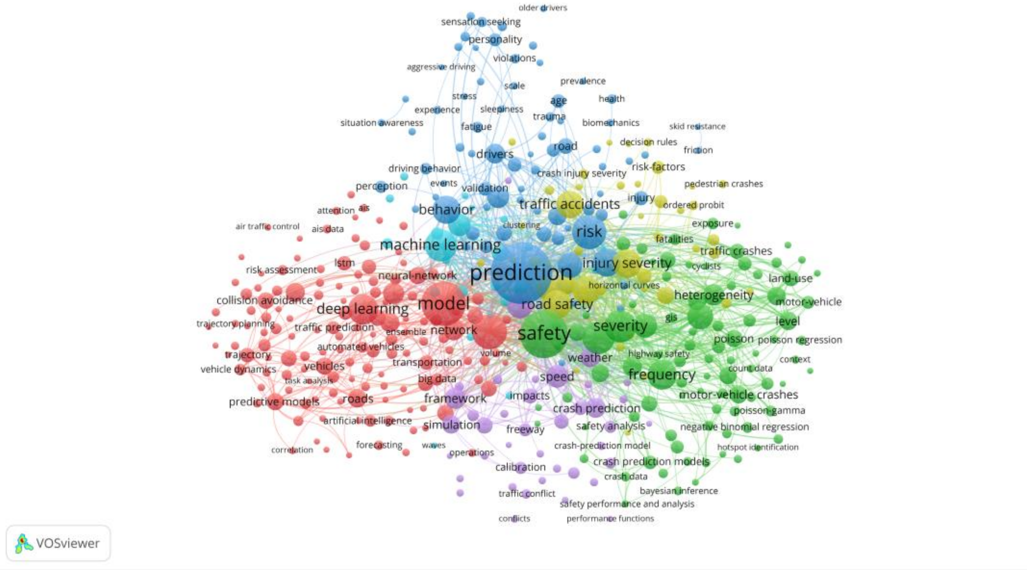

*Figure 1 – Traffic accident research network diagram*

### 1.5 Outline

The first section of this paper introduces the background and significance of traffic accident prediction, reviews the development history of traffic forecasting, and discusses the critical role of multi-source data in accident prediction. The second section provides a detailed description of the sources of multi-source data and their processing methods, with an in-depth discussion of traffic accident prediction methods and applications. The third section discusses the research results and the current research limitations.

## 2. RESEARCH MATERIALS AND METHODS

### 2.1 Research data sources

Traffic accident data comes from various sources, such as government datasets, open datasets, sensor data, social media data, geospatial data, private sector and crowdsourcing data [13]. Government and open datasets are crucial for accident prediction, offering comprehensive details on accidents, roads and landscapes, and forming a reliable foundation [21]. Sensor data from road equipment, in-car devices and smartphones provide precise traffic condition information [22]. Social media has emerged as a valuable source, offering supplementary crash prediction insights [23]. Geospatial data captures key information about road conditions, traffic flow and road topology [24]. Private sector data, such as that from public transport companies and traffic monitoring firms, along with crowdsourced data (e.g. reports of accidents and road conditions via smartphone apps), also play a significant role in supporting accident prediction [25, 26].

### *Government datasets and open datasets*

Government datasets, maintained by agencies like police and traffic departments, provide detailed, authoritative records of traffic accidents, driver and vehicle information, making them a crucial data source for traffic accident modelling [27]. Open datasets, freely accessible to the public, typically offer demographic, weather and road network data. Examples include the U.S., U.K. and Australian government open data catalogues [28]. Government datasets offer detailed and precise accident records, while open datasets expand the scope and depth of analysis. By comparing and integrating these two types of datasets, we can gain a more comprehensive understanding of traffic accident patterns, trends and potential risk factors.

Both government datasets and open datasets provide high reliability, large capacity and diverse information for accident prediction. However, data missing and outliers require extra handling [14]. It is necessary to quantify the data collected by different organisations and regions through standardised operations [29].

### *Sensor data*

The application of sensor technology enables access to extensive amounts of multi-source heterogeneous data related to traffic accidents, including vehicle GPS information, traffic status and weather data. Modern vehicles are commonly equipped with vehicle-mounted sensors, which can provide real-time information on vehicle conditions, driving behaviour and road environment [30]. Integrating deep learning and clustering algorithms facilitates real-time accident prediction [31] and the identification of high-risk areas [32].

Sensors collect precise, real-time traffic data, enabling the forecasting of accidents and helping emergency management departments optimise resource allocation and traffic control. However, real-time monitoring of the entire road network remains challenging due to the high costs and limited coverage of road sensors. Additionally, issues with data quality and integrity require costly preprocessing of large datasets [33].

### *Social media data*

In recent years, the widespread use of social media platforms like microblogging, Twitter and Facebook has made them valuable sources for traffic accident information. Social media data, which includes text, images, videos and voice recordings, has become an important channel for extracting details such as vehicle location, speed, accident type and severity [34]. This makes social media highly useful for traffic accident prediction, offering distinct advantages. For instance, Lu and Hao retrieved traffic and weather-related tweets, integrated spatiotemporal features with weather data and built a warning model by analysing traffic events [35]. Other researchers have used machine learning and deep learning models to predict real-time accidents based on Twitter traffic data [36].

Social media data offers several benefits for accident prediction, including low acquisition costs, high real-time accuracy and diverse information sources [16]. It serves as an effective supplement to traditional traffic data. However, user-generated content is prone to false, repetitive or irrelevant information, requiring rigorous data filtering to ensure prediction accuracy [37]. In addition, data containing user privacy information should be fully protected, and due to the geographical distribution of users, some regions may lack sufficient data for predictive analysis.

### *Geospatial data*

Geospatial information pertains to a defined geographic location on Earth's surface and offers insights into spatial distribution and attribute features. Consequently, it can be used to predict the location of traffic accidents. The Geographic Information System (GIS) analyses traffic accident data spatially and visually. It generates hotspots and heat maps to identify high-accident areas and potentially dangerous road sections by overlaying accident data with geographic elements [38, 39]. The Global Positioning System (GPS) employs satellite signals to detect and monitor mobile entities on the ground in real time, acquiring data such as traffic flow and accident sites. It can determine potential contributing factors to accidents by scrutinising the driving behaviour and trajectory leading up to the accident [40].

Due to the need for comprehensive transportation infrastructure, strong technological research and development, and data integration capabilities for the collection and application of geographic spatial data, the use of it for traffic accident prediction is mainly concentrated in technologically advanced countries such as China, the United States and Germany [41].

*Private sector and crowdsourcing data*

In recent years, the role of the private sector and crowdsourced data in traffic accident prediction has become increasingly important. Transportation service companies, bike-sharing and car-sharing companies collect a large volume of detailed data on vehicle locations, speeds and routes. Public transportation companies, equipped with advanced onboard monitoring systems, record real-time driver behaviour data, such as hard braking, sharp turns and speeding [42]. These high-quality and frequently collected data provide detailed insights into driver behaviour and its impact on traffic accidents [43]. Given that public vehicles cover a wide road network, especially in urban areas, their data offer comprehensive reflections of regional traffic conditions [44]. Bike-sharing companies also contribute real-time cycling data, which supplement traditional traffic data and help identify risky cycling behaviours [45]. Integrating these data with other datasets allows for a deeper exploration of causes and patterns in bicycle accidents, pedestrian collisions and multi-vehicle incidents [46]. Crowdsourcing platforms gather real-time data from a large number of users, including traffic flow, road conditions, weather changes, etc., which can reflect potential risk factors that may lead to accidents on the road in real-time. Compared to traditional fixed sensors or government data sources, crowdsourcing data can cover different areas more quickly and widely [47].

When using data from private enterprises and crowdsourcing platforms, it is essential to address concerns regarding data privacy, security, quality, accuracy and legality. Ensuring the integrity and reliability of the data throughout collection, processing and analysis is crucial for trustworthy predictive results.

### 2.2 Research data process

To effectively mine traffic accident features, research data must undergo appropriate preprocessing to improve quality and usability. This paper will provide a detailed introduction to data cleaning, feature extraction, standardisation, normalisation and data balancing [48]. Additionally, it will address the challenges of fragmented traffic accident data and present some effective solutions.

*Data cleaning*

The initial stage of data preprocessing is data cleaning, which involves handling missing values, removing noise and redundancy and correcting anomalies [49]. Empty values in traffic accident data can lead to incomplete datasets and loss of accident features. In order to improve the integrity and validity of the data, it is necessary to fill in or delete these values [50]. Repairing abnormal traffic accident data can improve data consistency and stability, and enhance the accuracy of model analysis and prediction. Duplicate information in the original data can reduce data diversity and exacerbate the impact of some features. Deleting or merging duplicate values can reduce data redundancy and noise [51].

When dealing with missing values, it is essential to choose suitable methods that consider the real circumstances to prevent excessive eliminations and the loss of a significant amount of valuable data [52]. When selecting and deleting duplicate data, it is important to manually label the information to be deleted. Labelling information for deletion in a large dataset is a challenging task [53].

*Feature extraction*

Feature extraction is crucial in traffic accident data processing. It can improve the predictive ability of the model, and reduce data dimensionality to enhance computational efficiency, while reducing noise interference, revealing potential accident-influencing factors, and enhancing the interpretability of the model. In addition, feature extraction helps to integrate multi-source data, and optimise the overall performance and prediction performance of the model.

Statistical methods like principal component analysis (PCA) [54], linear discriminant analysis (LDA) [55] and independent component analysis (ICA) [56] help extract class-specific and significant features from complex data. Geometric-based techniques, such as multi-dimensional scaling (MDS) and manifold learning, explore the data's geometrical structure to uncover latent features [57]. Image processing methods capture characteristics like textures, shapes, edges and colours [58]. Feature extraction needs to exclude features that are not relevant to the problem and reduce the effect of noise on the model.

*Standardisation and normalisation*

Traffic accident prediction typically requires data of different types and sources, but the significant differences in scale and magnitude between different types of data pose challenges for evaluating the relative importance of each feature. Standardisation or normalisation can effectively eliminate scale differences between different features, making traffic accident data more comparable [59]. Moreover, these techniques can accelerate model convergence, prevent feature bias and enhance data interpretability [60].

Common standardisation and normalisation methods comprise min-max normalisation and Z-score standardisation. Min-max normalisation maps the data into specified intervals. It preserves the original data structure, making it suitable for cases where the data are stable and free from outliers [61]. Standardisation by Z-score adjusts the data distribution to eliminate the impact of outliers or noise. This technique is particularly useful when the data are unevenly distributed or contain outliers [62].

*Data balancing*

The occurrence of traffic accidents exhibits significant temporal and spatial variability, leading to a notable imbalance between accident and non-accident samples [63, 64]. This imbalance adversely affects model learning, causing it to favour predicting the majority class while ignoring the impact of the minority class [65]. Data balancing is performed on traffic accident datasets to rectify the disproportion of positive and negative samples, enabling better identification of accident patterns and factors that influence their occurrence [66].

Resampling techniques, including oversampling, undersampling and hybrid sampling, are commonly used to balance datasets. Oversampling methods, such as SMOTE, duplicate or synthesise minority class samples through interpolation in feature space, increasing their representation [67]. Undersampling methods, like random undersampling and Tomek links, reduce the number of majority class samples to achieve balance [68, 69]. Hybrid sampling combines both approaches, augmenting minority samples while reducing majority ones [70]. Additionally, data generation techniques leverage artificial intelligence or machine learning to extract valuable information from original data and create new data [71]. Anomaly detection improves model performance by identifying and processing outliers, reducing dataset noise [72]. Zero-inflated models effectively address the issues of overdispersion and zero-inflation in count data, enabling more accurate modelling and prediction of sparse data [73]. Cost-sensitive learning further mitigates data imbalance by assigning different weights to samples during training [74]. Finally, it is essential to reassess model performance after data balancing, ensuring the independence of training and test sets to accurately reflect the actual data distribution.

*Data fragmentation processing*

The fragmentation of traffic accident data has emerged as a significant challenge in modern traffic management and safety analysis. These data originate from diverse sources, including police records, hospital reports, insurance data and traffic management systems, each employing different formats and standards. This heterogeneity hinders data integration and sharing, ultimately affecting the accuracy and effectiveness of traffic accident analysis. To address this, researchers are developing new data standards, interoperability frameworks and integration tools.

DATEX II is a European standard designed to facilitate the exchange and sharing of road traffic information and traffic management data. This standard encompasses various aspects, from data formats and models to exchange protocols, aiming to improve traffic data interoperability [75]. DATEX II supports real-time data exchange, enabling researchers to access the latest traffic flow information [76]. Similarly, the Traffic Management Data Dictionary (TMDD) standardises data exchange between traffic management systems, supporting communication among management centres for improved incident mitigation and event management [77]. Vehicular ad-hoc networks (VANET) is a network technology designed for communication between vehicles and between vehicles and infrastructure. It addresses fragmented traffic accident data by facilitating real-time communication between vehicles (V2V) and between vehicles and infrastructure (V2I) [78]. Due to network fragmentation during sparse traffic, timely notifications are hindered [79]. Therefore, researchers have enhanced VANET connections through roadside units (RSUs) to improve information transmission in high-speed scenarios [80]. While VANET focuses on local communication, the Internet of Vehicles (IoV) connects vehicles to cloud services for broader data exchange and analysis. IoV's real-time communication capabilities enable immediate monitoring and response to traffic accidents [81]. Some researchers have leveraged blockchain and machine learning to develop IoV edge servers, achieving low-latency transmission and a prediction accuracy of 90% [82].

### 2.3 Research methods

Predicting traffic accidents is crucial for proactive safety strategies and road safety improvement. Various predictive models have been developed and continuously optimised [83], generally classified into three categories: statistical learning methods [84], traditional machine learning methods [85] and deep learning methods [86]. The performance of these models varies depending on traffic data, road features, environmental factors and the specific problem being addressed [12]. Thus, a broad comparison of models is impractical, as performance may differ significantly under different conditions [13]. This paper introduces the characteristics, application scenarios and limitations of these methods in predicting traffic accidents using multi-source data.

*Statistical learning*

Statistical learning is the first technique to predict traffic accidents, grounded in assumptions about data distribution and supported by mathematical theory, focusing on parametric inference [87]. Historical data are analysed to explore the relationships between traffic accidents and factors like traffic flow, weather and road type, helping identify key variables and patterns [88]. The following section introduces several commonly used statistical learning methods for investigating traffic accident data with time dependencies, primary factors of accidents, accident counts, binary or multinomial types of accidents and ordered classification of accidents.

Traffic accident prediction holds significant importance for enhancing road safety and optimising traffic management. With the widespread application of time series data, the development of accurate accident prediction models has become a focal point of research. Among time series methods, the ARIMA model stands out for effectively capturing trends and seasonal variations in data [89]. However, its performance depends on factors like historical data richness, feature selection and parameter tuning. To address ARIMA's limitations, combining it with other methods can improve prediction accuracy. For example, Chen et al. combined ARIMA with MLR to predict traffic accident fatality rates using variables like traffic signs and lane areas [90]. Accidents are influenced by factors such as individual behaviour, vehicles, road conditions and the environment. Identifying these factors is essential for accident prevention and transport planning. Linear regression models help analyse key factors like traffic flow and road infrastructure [91] and are effective when the dependent variable is continuous. For discrete count data, such as accident frequency, traditional linear regression may not meet key assumptions like normal distribution and linearity [92]. In these cases, the negative binomial regression model is more suitable [93]. Frequently used for predicting accident frequency, it establishes correlations between independent variables and the number of accidents while accounting for the discrete nature of the data [94].

Forecasting road accident severity is critical for traffic safety management, risk assessment, emergency response and transport planning, helping reduce accidents and mitigate damages. Logistic regression, a commonly used linear model for binary classification problems [95], maps the output of linear regression to a probability space via a logistic function [96]. Variables like weather, traffic flow and vehicle type are transformed into the numerical form using one-hot encoding or standardisation [97], with the sigmoid function mapping prediction results between 0 and 1 to indicate whether an accident will occur [98]. For predicting multiple levels of accident severity, the ordered probit model is more appropriate. It not only predicts categorical outcomes but also estimates the ordering of dependent variables across categories [99]. For instance, accident severity can be classified from minor to severe [100], and the probability of transitions between these levels can be determined based on various independent variables [101].

The statistical learning methods discussed each has their unique strengths but also come with limitations. Linear regression struggles to fit data accurately due to the probabilistic nature of traffic accidents and the complex nonlinear relationships in multi-source data [102]. Negative binomial regression relies on specific assumptions and is sensitive to outliers and overdispersion, which can lead to endogeneity issues and skew parameter estimates [103]. Logistic regression, when faced with large feature spaces or numerous multi-class variables, often underfits the data [104]. The ordered probit model requires strict assumptions, such as independent observations and no multicollinearity, which can limit its application [105].

*Table 1 – Analysis of the literature on the prediction of road traffic accidents based on statistical learning algorithms*

| Author | Data source region | Real-time detection | Dataset type | Accident information | Road information | Vehicle information | Weather information | Other information | Model | Evaluation index |
|---|---|---|---|---|---|---|---|---|---|---|
| [89] | Amhara | × | A | √ | × | × | × | × | ARIMA | Standard deviation |
| [90] | Shanghai | × | A, B | √ | × | × | × | √ | ARIMA, MLR | MAE, RMSE |
| [106] | London | × | A, B | √ | × | √ | √ | × | ARIMA, SARIMAX | MSE, RMSE, MAE, MAPE |
| [91] | Ireland | × | A, B | √ | √ | √ | √ | √ | GLM | AUC、ROC、CI |
| [107] | Chile | × | A, D | √ | × | √ | √ | × | GLM | Speed, density |
| [92] | Hong Kong | × | A, B | √ | √ | × | × | × | NBR | AIC |
| [94] | Singapore | × | A, C | √ | √ | × | × | × | RENB | Ratio of log-likelihood index ( $\rho^2$ ), ratio of log-likelihood ( $R^2$ ) |
| [108] | Turkey | × | A, B | √ | × | × | × | × | NBR, Poisson regression, empirical Bayesian | Z-score, standard error |
| [25] | Melbourne | × | A, B, F | √ | √ | × | × | √ | Mixed logit | Standard deviation |
| [44] | Mashhad | × | F | √ | √ | × | × | √ | Binary logit | Chi-square test, OR |
| [84] | Florida | √ | A, B, E | √ | √ | × | √ | × | Multilevel Bayesian logistic regression | DIC, AUC, accuracy |
| [95] | Eskisehir | × | A, B | √ | √ | √ | × | √ | Logistic regression, discriminant analysis | Probability value, OR |
| [96] | Nanjing | × | Drone data | √ | √ | × | × | × | Multilevel logistic regression model | TTC |
| [98] | Ghana | × | A, B | √ | √ | √ | × | √ | Ordered logistic regression | Rates between different types of traffic accidents |
| [99] | Washington | × | A, B | √ | √ | √ | √ | √ | Correlated random parameters ordered probit | Likelihood ratio, AIC |
| [100] | Hokkaido | × | A, B, D, F | √ | √ | × | √ | × | Ordered probit | Standard error, coefficient, t-value, Chi-square test |
| [101] | Florida | × | A, B | √ | √ | √ | × | √ | Ordered probit | Likelihood ratio |

** A, B, C, D, E and F represent government datasets, open datasets, sensor data, social media data, geospatial data and private sector and crowdsourcing data, respectively.*

As shown in *Table 1*, this section provides a detailed compendium of the literature referred to by the statistical learning methods described above. A comprehensive analysis of statistical learning methods yields the following conclusions. First, most data used in the literature come from government agencies and public sources, as these datasets are comprehensive and accurate, providing a reliable foundation for model development. Second, statistical learning algorithms, grounded in mathematical theory, are effective in identifying linear relationships between variables and traffic accidents. However, the inherently random and disordered nature of accidents makes it difficult to fully explain them using purely mathematical approaches. Although statistical learning has gradually been replaced by machine learning and deep learning models with stronger self-learning capabilities when using big data and multi-source data to predict accidents [109], statistical learning methods still exhibit more stable and superior performance in situations where the dataset is small and the number of features is limited [13].

### *Machine learning*

Machine learning provides a clear advantage over statistical learning in predicting traffic accidents. It can handle vast amounts of data to learn patterns, identify data features and laws and reveal complex, nonlinear correlations in traffic data [110]. The adaptive learning and adjustment capabilities allow machine learning to update models and parameters in response to new data [111]. Supervised learning is often used to predict accident severity by classifying or regressing based on known labels such as time, location, weather, vehicle type and driver age. This paper explores common supervised learning methods, including decision tree (DT), random forest (RF), support vector machine (SVM) and naive Bayes (NB). It provides an analysis of their applications, strengths and limitations in traffic accident prediction.

DTs have garnered attention for their ability to represent decision rules and predictions in a tree structure, making them effective for assessing the impact of key features [112]. They are particularly suitable for predicting accidents influenced by various factors, and pruning techniques can reduce overfitting by removing less significant variables [113]. However, DT models may suffer from randomness and fail to reliably rank the importance of variables. RF addresses this by building multiple decision trees, averaging the significance of each feature across trees, and improving prediction accuracy and robustness [114]. DT and RF utilise tree structures to classify various categories. In contrast, SVM and NB offer distinct classification approaches. SVM maximises the margin between classes by mapping data into a high-dimensional feature space and seeking the optimal hyperplane for separating categories [115]. It performs well in binary classification, such as identifying accident-prone patterns in real-time [116], but its performance may decline with larger datasets and more classification categories [117]. NB calculates posterior probabilities based on feature vectors and assigns inputs to categories such as accident severity [118]. While NB is efficient in handling large, high-dimensional datasets, its assumption of feature independence can limit accuracy in complex traffic environments [119, 120]. Combining it with other algorithms or feature engineering may improve performance.

Unsupervised learning uses unlabelled data to uncover hidden structures or patterns. This approach can be used to cluster accident data based on unknown characteristics or labels, such as driving behaviour, road conditions and accident causes. Common unsupervised models include fuzzy C-means (FCM), K-means and DBSCAN. These clustering algorithms help categorise different types of accidents or risk levels by analysing patterns and relationships within the data. Clustering algorithms have the ability to categorise varying types of accidents or levels of risk by analysing connections, structure and patterns among datasets. K-means groups data points based on the Euclidean distance to the nearest cluster centre but requires a predetermined number of clusters [121]. As Kumar’s research has shown, this method is efficient and interpretable, and they used threshold-based clustering to classify accident-prone locations [122]. However, K-means assumes each point belongs to only one cluster, which can oversimplify real-world scenarios [123]. In contrast, FCM allows data points to belong to multiple clusters by assigning membership degrees. This flexibility makes FCM well-suited for handling uncertainty in traffic accident data and identifying factors contributing to accident risk [124]. Both the K-means and FCM require the number of clusters (K) to be specified in advance, which complicates finding the optimal K for model performance. DBSCAN is a density-based clustering algorithm that identifies varied accident patterns and pinpoints anomalies or accident hotspot regions by segregating accident data into high and low-density clusters [125]. Additionally, DBSCAN demonstrates high computational efficiency and rapid clustering speed. It can identify clusters of any shape and define their number and boundaries based on density [126]. Nevertheless, it exhibits poor clustering performance when the data density is uneven, the cluster spacing is considerable or the data are large in dimension.

*Table 2 – Literature analysis of road traffic accident prediction based on machine learning*

| Author | Data source region | Real-time detection | Dataset type | Accident information | Road information | Vehicle information | Weather information | Other information | Model | Evaluation index |
|---|---|---|---|---|---|---|---|---|---|---|
| [42] | Chongqing | × | F | √ | × | √ | × | √ | Boosted trees | Precision, recall, F1, FPR |
| [45] | Chengdu | × | B, F | × | √ | × | × | √ | NBADT | Relative importance |
| [110] | Granada | × | A | √ | √ | √ | √ | √ | DT | IGR, Ginf |
| [112] | Saskatchewan | × | A | √ | √ | √ | √ | × | ID3, C4.5 | Correctly classified instances, incorrectly classified instances |
| [113] | Zhongshan | × | A, B | √ | √ | × | × | √ | GBDT | MAPE |
| [85] | America | × | A, B | √ | × | √ | × | √ | RF | Out-of-bag error rate |
| [111] | Malaysia | × | A, B, F | √ | √ | √ | × | √ | CART, RF | TPR, FPR, precision |
| [114] | Harbin | × | A | √ | × | √ | √ | × | RF, LightGBM | ROC, AUC, accuracy |
| [16] | Italy | √ | D, F | √ | × | × | × | √ | SVM | Accuracy, F1-score, recall precision |
| [26] | California | √ | A, B, F | √ | √ | × | √ | × | SVM | Precision, recall |
| [127] | Wuhan | × | A | √ | √ | × | √ | × | SVM | Accuracy, recall, F1-score, AUC |
| [17] | Pittsburgh, Philadelphia | √ | D, F | √ | × | × | × | √ | SNB | Accuracy, recall, precision |
| [118] | California | × | A, B | √ | × | × | × | × | NB | Recall, precision |
| [121] | Hungary | × | A, B | √ | × | × | × | × | K-means | Black dot ratio |
| [123] | London | × | A, B | √ | √ | × | √ | × | K-means | Variance |
| [128] | Teheran | × | C | × | × | √ | × | × | K-means | Correct TCR percentage |
| [124] | America | × | A, B | √ | × | √ | × | √ | Fuzzy clustering, DT | Fit scores |
| [129] | Medan | × | A, B | √ | × | × | × | × | FCM | Consistency level |
| [125] | Hunan | × | A | √ | √ | √ | √ | √ | DBSCAN, BN | Prior probability, posterior probability |
| [126] | China | × | A | √ | × | × | × | × | AD-DBSCAN | Calinski-Harabasz index |
| [130] | Portugal | × | A | √ | √ | × | √ | × | KDE, DBSCAN | Moran-I |

** A, B, C, D, E and F represent government datasets, open datasets, sensor data, social media data, geospatial data and private sector and crowdsourcing data, respectively.*

This section provides a detailed review of the machine learning methods discussed in the literature, summarised in *Table 2*. The following conclusions can be drawn from the summary of machine learning applications in predicting traffic accidents. Machine learning is broadening the scope of datasets used for accident prediction by incorporating data from social media and sensors for real-time detection. This capability is due to machine learning's robust data processing abilities, which can analyse high-dimensional data and capture complex nonlinear relationships [20]. Therefore, utilising multi-source datasets may improve the effectiveness of machine learning. Finally, while machine learning has advanced considerably compared to statistical learning for various application scenarios and multi-source data utilisations, it necessitates significant computational resources, demands high-quality data and may involve processes like feature engineering that heighten the expenses of model computation [28].

### *Deep learning*

Deep learning technology has brought revolutionary changes to traffic accident prediction in recent years. As a machine learning technique using artificial neural networks, deep learning has significantly improved the performance and application of traffic accident prediction. It autonomously learns complex accident patterns through multi-level nonlinear transformations.

A large amount of image and video data is collected during vehicle operation via traffic monitoring systems and onboard cameras. Analysing these data helps identify the vehicle's movements before an accident and the influence of environmental factors [131]. Convolutional neural network (CNN) can extract significant features from images through convolutional and pooling layers. It can learn essential characteristics of accident occurrence and spatial data correlation [132, 133]. CNN has been successfully applied to real-time detection of potential accidents using sensor data from transportation systems [134]. The input data for CNN are usually Euclidean data, such as images. To ensure computational efficiency and model performance, the usual approach is to resize the image to a uniform size while maintaining a regular structure. However, CNN can be prone to overfitting, particularly with limited training samples. Regularisation techniques or dropout are commonly used to mitigate this issue and enhance generalisation [135]. Graph convolutional networks (GCN), introduced by Thomas N. Kipf and Max Welling in 2017, have emerged as effective tools for handling non-Euclidean data, such as graph or manifold data. GCN learns node features by performing convolution operations on graph-structured data [136]. In traffic networks, GCN can process irregular graph structures like road networks, accurately capturing spatial correlations between traffic nodes. This allows for effective detection and analysis of traffic accidents across the network [137]. However, preprocessing adjacency matrices to represent the connectivity of road networks requires specialised knowledge in the field, and GCN is sensitive to data noise.

Both GCN and CNN are effective in capturing spatial dependencies in data, while recurrent neural networks (RNN) are adept at learning temporal dependencies. RNNs are particularly useful for analysing sequential data, and capturing timing and semantic relationships [138]. It can better understand the relationship between short-term data, but it is difficult to learn information in long-term time series. Long short-term memory (LSTM) is a specialised type of RNN that introduces cell states and gating mechanisms, allowing it to retain relevant information over extended periods and learn both short-term and long-term dependencies in traffic accident data [139]. Additionally, LSTM can be trained via backpropagation algorithms, eliminating the need for manual tuning and mitigating gradient explosion issues seen in RNN [19]. Multiple LSTM models can be applied to accommodate time series data of varying lengths, predicting traffic accident risks at different levels of granularity and periodicity (such as daily, weekly or monthly) [140]. However, LSTM has high data quality requirements and is prone to overfitting. Data augmentation or regularisation techniques can be used to avoid the above problems.

Traffic collision prediction is complex due to factors like road network topology, temporal traffic flow changes and multi-source data fusion [24]. Therefore, achieving the desired accuracy to predict traffic accidents is challenging when relying solely on one model. Consequently, it is necessary to adopt an effective strategy of integrating various deep learning models for accident forecasting, which utilises the strengths of individual models and enhances prediction performance [141]. For example, the combination of CNN and LSTM algorithms simultaneously considers the spatial and temporal characteristics of traffic accidents, enhancing their predictive accuracy and comprehensiveness [142]. Integrating deep learning algorithms necessitates careful consideration of data consistency and compatibility, guaranteeing harmonious data inputs and outputs across multiple algorithms.

*Table 3 – Literature analysis of road traffic accident prediction based on deep learning*

| Author | Data source region | Real-time detection | Dataset type | Accident information | Road information | Vehicle information | Weather information | Other information | Model | Evaluation index |
|---|---|---|---|---|---|---|---|---|---|---|
| [37] | America | √ | D, F | √ | × | × | × | × | CNN, RNN | Accuracy, F1-score precision, recall |
| [131] |  | √ | B, D | √ | × | × | × | √ | CNN | Accuracy, precision, F1-score, recall |
| [132] | UK | √ | A | √ | × | √ | √ | × | CNN, RF | MSE, AUC |
| [133] | Madrid | × | A, B | √ | × | √ | √ | √ | CNN | Recall, F1-score, precision |
| [134] | Des Moines | √ | A, B, C | √ | × | √ | × | × | CNN, ANN | ROC-PR, F1-score |
| [143] | Nashville | √ | A, B, F | √ | × | × | √ | √ | CNN | F1 score, average early prediction ratio, average early prediction distance, average early prediction time |
| [144] |  | × | Simulator data | √ | × | × | × | × | DCNN | Accuracy, precision, TPR, FPR |
| [109] | Beijing | √ | A, B, F | √ | √ | √ | √ | √ | DSTGCN | RMSE, recall, F1-score precision, AUC |
| [136] | San Diego, Los Angeles | × | A, B, C | × | × | √ | × | √ | GCN, LSTM | Precision, F1-score, recall AUC |
| [137] | New York, Chicago | × | A, B, F | √ | √ | √ | √ | √ | GCN, CNN | RMSE, MAP, accuracy |
| [145] | New York, California | × | A, B | √ | √ | √ | √ | × | MADGCN | Recall, precision, F1-score, AUC |
| [3] | UK | × | A, B | √ | × | × | × | √ | LSTM-GBRT | RMSE, R-square, RMSLE |
| [19] | California | × | A, B | √ | × | √ | × | √ | LSTMDTR | Accuracy, F1-score, recall, AUC, precision |
| [140] | China | × | A | √ | × | × | × | √ | LSTM | MAE, MSE, RMSE |
| [23] | Northern Virginia, New York | √ | D, F | √ | × | × | × | √ | LSTM, DBN | Accuracy, precision |
| [24] | New York | × | A, B, C, F | √ | √ |  | √ | √ | CNN, LSTM | MSE, MAE, MAPE |
| [86] | Florida | × | A, B, C | √ | × |  | √ | √ | CNN, LSTM | AUC, TPR, FPR |
| [139] | California | × | A, B, F | √ | × |  | × | √ | LDA, LSTM Bi-LSTM | RMSE, MAE, MAPE |
| [141] | Taiwan | × | A | √ | √ |  | × | × | ML, CNN, DNN, DBN | Accuracy, F1-score, recall, precision |
| [142] | Ningbo | × | A, B | √ | × |  | × | √ | Bi-ConvLSTM, U-Net | CE, MSE, RMSE, CSI, FAR, POD |
| [146] | Paris | × | A, B | √ | × |  | × | √ | LSTM, CNN, ANN | Accuracy, recall, FAR, ROC |

* *A, B, C, D, E and F represent government datasets, open datasets, sensor data, social media data, geospatial data and private sector and crowdsourcing data, respectively.*

This section offers a comprehensive review of the literature referenced in the previous deep learning techniques, illustrated in *Table 3*. The following findings are obtained by summarising the application of deep learning in predicting traffic accidents. First, the intricate configuration of deep learning models typically necessitates establishing several crucial structural layers, such as the convolutional layer and the pooling layer in CNN. Furthermore, the numerous parameters involved require extensive data to support model training and parameter adjustment. Hence, deep learning can capitalise on the benefits of multi-source datasets to a greater extent [147]. Second, the multilevel constructs and advanced data processing capabilities of deep learning permit it to acquire intricate traffic accident patterns and features while handling nonlinear, high-dimensional and unstructured information [109]. Deep learning allows highly parallelised computing using GPUs to accelerate model training and inference. Finally, the integration of algorithms that combine various models can utilise their respective strengths to capture more comprehensive features to predict accidents.

*Pioneering research*

The above content systematically introduces the applications of statistical learning, machine learning and deep learning in traffic accident prediction, and analyses the characteristics of different models. However, existing research cannot cover all methods, and the following are some groundbreaking and inspiring studies.

Lv et al. first attempted to apply the K-nearest neighbour (KNN) algorithm to the real-time prediction of highway accidents. Based on features such as vehicle speed, flow rate and density, it was able to predict 80% of dangerous traffic conditions, demonstrating the advantages of the algorithm's simple structure and adaptability to nonlinear relationships [148]. With the rapid development of transportation big data, the predictive performance of traditional shallow learning models is gradually being questioned. Lv et al. used a stacked autoencoder (SAE) in deep learning for traffic flow prediction, successfully matching traffic patterns under large and medium traffic volumes, providing a new direction for future traffic accident prediction [149]. In the same year, Ma et al. extended deep learning theory to large-scale network analysis, promoting research on traffic network congestion and accident prediction [150]. In the field of predicting the severity of traffic accidents, Yang et al. proposed the first multi-task deep neural network framework that can predict different degrees of injury, death and property damage, and improve the interpretability of the model through hierarchical correlation propagation [151]. The multi-task prediction model and model interpretability provided in this study have important implications for improving the prediction range and interpretability of other models. Reinforcement learning can provide more accurate accident prediction and decision support in complex and dynamic traffic environments by continuously interacting and optimising strategies. Cho et al. first applied double actors and regularised critics (DARC) to traffic accident prediction and significantly improved the safety of autonomous driving by using driving recorder videos as input data [152]. The method based on macroscopic road network images proposed by Ji et al. eliminates the dependence on detailed traffic dynamics and data, providing new ideas for accident prediction in situations where data are insufficient [153].

### 2.4 Model evaluation

In the study of traffic accident prediction, the evaluation of model performance is a critical component. By quantifying the accuracy and effectiveness of predictive models, we can effectively distinguish the strengths and weaknesses of different models, ensuring the reliability of the prediction results. This process also provides a scientific basis for the optimisation and adjustment of models. In this section, we will provide an in-depth introduction to the commonly used evaluation metrics for both regression and classification problems, aiding in a better understanding of model performance.

*Regression problem*

Regression problems involve predicting continuous numerical variables. In traffic accident prediction, common regression problems include traffic flow prediction, accident quantity prediction and accident severity prediction. The commonly used evaluation indicators include mean square error (MSE), which is the average square of the difference between predicted and actual values. It measures the degree of dispersion between predicted and actual values, and the smaller the better. Root mean square error (RMSE), which is the square root of MSE, ensures that the unit of error is consistent with the original data, making it easy to interpret. Mean absolute error (MAE) refers to the average absolute difference between predicted and actual values, without considering the direction of the error, only focusing on its magnitude. The mean absolute percentage error (MAPE), which is the percentage average of the absolute error between the predicted value and the actual value, represents the relative magnitude of the error, facilitating comparison between data of different dimensions.

*Classification problem*

Classification problems involve predicting discrete category labels, with common tasks in traffic accident prediction including determining the likelihood of an accident occurring and forecasting the type of accident. Evaluation metrics for these classification tasks typically encompass accuracy, precision, recall, F1 score, receiver operating characteristic (ROC) curve and area under the curve (AUC). Accuracy measures the ratio of correctly predicted samples to the total number of samples, reflecting the overall classification effectiveness of the model. Precision indicates the proportion of true positive cases among samples predicted as positive, focusing on the reliability of positive predictions. Recall assesses the proportion of actual positive cases that are accurately identified. The F1 score, which is the harmonic mean of precision and recall, ranges from 0 to 1, with values closer to 1 denoting superior model performance. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold levels, while AUC quantifies the area under the ROC curve, with values approaching 1 indicating enhanced model performance.

## 3. RESEARCH OF PREVIOUS STUDIES AND DISCUSSION

### 3.1 Results of previous studies

1) Detailed traffic accident data can significantly enhance the accuracy and generalisation ability of predictive models. This paper reviews various data sources, including government and open datasets, sensor data, social media data, geospatial data, private sector and crowdsourcing data. First, government and open datasets provide comprehensive and reliable records of traffic accidents, road conditions and environmental data. However, due to differences in data collection and management methods across regions and institutions, data quality and storage structures are often inconsistent, requiring rigorous preprocessing to ensure uniformity. Second, while sensor data are costly and have limited coverage, their high precision and real-time nature are crucial for real-time accident prediction. Social media data, in the form of text, images and other media, offer a low-cost source of traffic-related information. Despite the presence of false, redundant or irrelevant content, effective filtering can make these data a valuable supplement to predictive models. Additionally, geospatial data, obtained through GIS and GPS technologies, reveal the spatial distribution of traffic accidents, aiding in the identification of accident hotspots and hazardous road segments. However, their acquisition and processing can be expensive and require specialised analytical skills. Finally, private sector and crowdsourced data are collected through various traffic applications and platforms, yielding a wealth of real-time traffic information and user feedback. These data sources are diverse and have broad coverage, providing fine-grained traffic flow and road condition information. However, due to the varied methods of data collection, their quality and accuracy may be inconsistent, necessitating effective data cleaning and integration to ensure their effectiveness in predictive models.
2) Analysis of the data sources in the reviewed literature reveals that a substantial portion of research data originate from developed countries such as those in Europe and North America. This trend can be attributed to several factors. Firstly, these countries have well-established traffic management systems and data collection mechanisms, characterised by high-quality data recording and management systems, and are supported by substantial funding and resources for large-scale data collection. Additionally, research institutions and universities in these countries have extensive experience and influence in the field of traffic accident research, resulting in traffic accident data with high reliability and consistency. Particularly in countries like the United Kingdom and the United States, regulations such as the Freedom of Information Act and the Public Records Act mandate the public availability and access to government data, thereby promoting data transparency and facilitating the availability of traffic accident-related datasets to the public. Consequently, these datasets are frequently used as primary sources by researchers. In contrast, other regions face limitations in data accessibility and sharing due to stringent data privacy regulations, lower standardisation and issues related to data privacy and sharing.
3) Existing research indicates that data preprocessing is crucial for enhancing the quality and usability of traffic accident data. Specific methods include data cleaning, feature extraction, standardisation or normalisation and data balancing. Data cleaning aims to improve data integrity and consistency by

removing missing values, eliminating noise and redundancy, and correcting anomalies and errors. Feature extraction identifies key factors influencing accident occurrence and severity through various statistical, geometric and image processing techniques. Standardisation and normalisation eliminate scale differences between features, making the data more comparable. Addressing class imbalance issues through resampling techniques such as oversampling, undersampling and hybrid sampling, as well as data generation techniques like SMOTE, enhances the model's predictive capability for minority class samples. Additionally, to tackle the fragmentation of traffic accident data, existing research proposes establishing unified data collection standards and utilising methods such as VANET and IoV to achieve data collection, sharing and interoperability, thereby improving data completeness and consistency.

4) Traffic accident prediction is a critical challenge in the field of traffic safety, and researchers have developed various prediction models, mainly categorised into statistical learning methods, traditional machine learning methods and deep learning methods. Statistical learning methods such as linear regression, negative binomial regression, logistic regression and ordered probit models rely on historical data to identify key factors and their impacts on accident occurrence. However, they may be limited in handling complex nonlinear relationships. Machine learning methods such as DT, RF, SVM, NB, K-means, fuzzy C-means and DBSCAN classifiers leverage large datasets for training, enabling them to learn complex associations and nonlinear relationships within data. These methods adaptively adjust model parameters and are suitable for classifying or clustering traffic accident data. Deep learning methods such as CNN, GCN and LSTM employ multi-layer nonlinear transformations to autonomously learn complex patterns in traffic accident data. They excel particularly in handling graph data and time series data. Combining approaches from different deep learning models integrates their respective strengths, enhancing prediction accuracy and comprehensiveness. Therefore, different models exhibit their advantages and limitations in utilising multi-source data for traffic accident prediction. The choice of appropriate models should depend on specific application scenarios and data characteristics.

### 3.2 Discussion

Current research on predicting road traffic accidents based on multi-source data faces several challenges.

1) Despite the increasing richness of sources for traffic accident data, individual researchers still face significant information barriers. They typically rely on government datasets or open data to obtain information, while access to highly private and sensitive data remains challenging. Additionally, individual researchers find it difficult to obtain more detailed data from private enterprises (such as public transportation companies, insurance firms, bike-sharing or car-sharing companies) due to a lack of relevant support and resources. Therefore, acquiring multi-source data related to traffic accidents remains a significant challenge for individual researchers.
2) When predicting accidents based on multi-source data, model development typically requires integrating various types of datasets, such as historical accident data, weather data, road data, points of interest (POI) data and traffic flow information. This diversity in data sources presents challenges related to high-dimensional features and feature selection. Each dataset contains multiple features, which requires greater computational resources for feature processing. Additionally, there may be high correlations or redundant information among some features, which can reduce the efficiency of model training and even impact prediction performance. To address these issues, appropriate feature selection methods need to be employed. However, feature selection relies not only on effective algorithms but also on domain-specific knowledge. A lack of relevant domain knowledge may result in feature selection outcomes that do not accurately reflect the actual research context. Therefore, in feature selection, it is essential to use suitable methods in conjunction with the domain expertise to ensure the scientific rigour and effectiveness of the process.
3) In machine learning and deep learning, underfitting and overfitting are common challenges that affect model performance, especially in complex traffic patterns. Underfitting occurs when a model lacks complexity, failing to capture underlying patterns, leading to high bias and low accuracy [154]. Conversely, overfitting arises when the model is too complex and thus learns the patterns in the training data excessively, resulting in high variance and poor generalisation [155]. To address these issues, various techniques can be applied, including regularisation, ensemble learning, data augmentation, early stopping, cross-validation and dropout.

Regularisation, such as L1 (Lasso) and L2 (Ridge), penalises large weights to reduce model complexity and improve generalisation [156]. Ensemble learning combines multiple base learners to enhance overall model accuracy and generalisation, integrating weaker models into a stronger one [157]. In neural networks, large models can be built by combining smaller ones, acting as base learners [158]. Given the imbalance in traffic accident data, particularly the prevalence of zero inflation [156], data augmentation expands the training set by generating new samples through random transformations (such as rotation, scaling and flipping) [159], thus improving robustness and reducing overfitting. The early stopping mechanism is an effective method to prevent model overfitting by monitoring the performance of the model on the validation set (such as loss and accuracy) during the training process. When the performance no longer improves, the training is stopped in advance [160]. Cross-validation effectively reduces the risk of overfitting by dividing the dataset into multiple subsets and alternating training and testing on each subset. [161]. Dropout is a simple and effective technique that randomly disables neural network nodes during training, allowing the remaining nodes to compute forward and backward propagation, thereby improving generalisation [162].

## 4. CONCLUSION

### 4.1 Conclusion

This paper presents a systematic review of research on traffic accident prediction based on multi-source data. Through detailed analysis of data sources, data processing methods and predictive models, the following key conclusions are summarised:

1) The importance of multi-source data: Detailed traffic accident data significantly enhance the prediction accuracy and generalisation capability of models. Government data, sensor data, social media data, geospatial data, private sector and crowdsourcing data each have their advantages and limitations. Integrating these data sources can provide comprehensive and rich information support for models.
2) The necessity of data preprocessing: Effective data preprocessing is crucial to ensuring high-quality data input. Through data cleaning, feature extraction, standardisation or normalisation and data balancing, the quality and usability of the data can be significantly improved, thereby enhancing the performance and interpretability of predictive models.
3) Diversity of prediction methods: Traffic accident prediction methods include statistical learning methods, traditional machine learning methods and deep learning methods. Statistical learning methods are suitable for analysing the main factors contributing to accidents, traditional machine learning methods can handle complex nonlinear relationships, and deep learning methods perform exceptionally well on large-scale datasets. The appropriate prediction method should be chosen based on the specific application scenario and characteristics.

### 4.2 Future works

Predicting traffic accidents based on multi-source data involves specific research questions, innovative methods and potential directions for interdisciplinary collaboration, which will be introduced in the following sections.

*Establishing a unified vehicular networking data-sharing platform for predicting hazardous driving behaviour*

The IoV holds significant potential for traffic accident prediction by providing real-time, high-precision data on vehicle operations, such as location, speed and acceleration [163]. IoV data also capture driving behaviours like rapid acceleration, hard braking and sharp turns, which can be analysed to predict potential accident risks. Furthermore, IoV enables real-time information sharing between vehicles, alerting drivers to traffic incidents and road conditions, and facilitating timely assistance for accident victims. However, IoV data are underutilised in traffic accident prediction due to limited data-sharing mechanisms, with vehicle manufacturers and service providers maintaining tight control over the data and the absence of comprehensive, open data-sharing platforms.

To address these challenges, several measures can be implemented. (1) Develop standardised V2X data-sharing protocols to ensure interoperability between manufacturers and service providers, enabling secure data exchange. (2) Utilise cloud and edge computing for real-time analysis of large-scale V2X data, enabling timely accident predictions and warnings. (3) Design algorithms to recognise risky driving behaviours and issue alerts.

(4) Employ privacy-preserving techniques like differential privacy and federated learning to protect data during sharing and analysis. Achieving these solutions requires interdisciplinary collaboration with traffic authorities, vehicle manufacturers and cybersecurity experts to establish policies, acquire data, and ensure security and privacy, thus driving the advancement of connected vehicle technology.

### *Integrating and processing real-time information from various data sources to enhance traffic accident prediction accuracy*

Real-time traffic accident prediction can identify hazards and anomalies in traffic flow, enabling preventive measures such as adjusting traffic signals and alerting drivers, thereby reducing accidents [164]. This capability also assists emergency responders in quickly understanding the location, severity and potential impact of accidents, facilitating more effective rescue operations and traffic management. Currently, the excellent performance of deep learning plays a significant role in real-time traffic accident prediction.

To improve accuracy, several measures can be implemented. (1) Develop multi-source data fusion techniques to integrate data from social media, V2X, weather and road conditions for comprehensive monitoring. (2) Optimise deep learning models for real-time traffic accident prediction. (3) Build intelligent traffic systems to dynamically adjust signals and provide hazard alerts. (4) Establish emergency response systems to efficiently allocate resources based on prediction outputs. Achieving these requires collaboration with social media platforms, connected vehicle providers and emergency response departments to enhance data acquisition, predictive accuracy and rescue efficiency.

### *Constructing graph-structured data suitable for traffic accident prediction and enhancing GCN model performance*

GCN models exhibit significant advantages in handling non-structured data, particularly in analysing traffic accident risks across an entire road network using graph-structured data. However, the application of GCN models in the field of traffic accident prediction is currently limited.

To enhance the performance of GCN models in traffic accident prediction, several strategies can be implemented. (1) Design effective methods to transform traffic networks, traffic flow and accident data into graph-structured formats suitable for GCN models. (2) Improve GCN models by incorporating techniques like attention mechanisms and graph generative adversarial networks (GAN) to better capture complex traffic patterns. (3) Apply multi-task learning by integrating related tasks such as traffic flow prediction and congestion detection to improve model generalisation and accuracy. Achieving these advancements requires interdisciplinary collaboration with traffic planning experts to optimise graph data construction and model design, and with computer scientists to refine GCN algorithms, ensuring practical application and improved model performance.

## ACKNOWLEDGEMENTS

This research was funded by the Opening Project of the Intelligent Policing Key Laboratory of Sichuan Province (grant number No. ZNJW2023KFMS004).

## DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

## REFERENCES

[1] WHO.Who kicks off a decade of action for road safety.2021.https://www.who.int/news/item/28-10-2021-who-kicks-off-a-decade-of-action-for-road-safety

[2] WHO.Global status report on road safety 2023.2023.https://www.who.int/publications/i/item/9789240086517

[3] Zhang Z, et al. Traffic accident prediction based on lstm-gbrt model. J. Control Sci. Eng. 2020;2020:4206919. DOI:10.1155/2020/4206919.

[4] Trirat P, Yoon S, Lee J. Mg-tar: Multi-view graph convolutional networks for traffic accident risk prediction. IEEE Trans. Intell. Transp. Syst. 2023;24(4):3779-3794. DOI:10.1109/TITS.2023.3237072.

[5] Xiong X, et al. Multi-level prediction framework of driving risk based on the matter-element extension model. Transp. Res. Rec. 2024;2678(8):950-965. DOI:10.1177/03611981231223750.

[6] Liu X, et al. Attention based spatio-temporal graph convolutional network with focal loss for crash risk evaluation on urban road traffic network based on multi-source risks. Accid. Anal. Prev. 2023;192:107262. DOI:10.1016/j.aap.2023.107262.

[7] Pilutti T, Ulsoy A G. Fuzzy-logic-based virtual rumble strip for road departure warning systems. IEEE Trans. Intell. Transp. Syst. 2003;4(1):1-12. DOI:10.1109/TITS.2003.811810.

[8] Tan D, Chen W, Wang H. On the use of monte-carlo simulation and deep fourier neural network in lane departure warning. IEEE Intell. Transp. Syst. Mag. 2018;10(1):168-168. DOI:10.1109/MITS.2017.2776726.

[9] Qu X B, et al. Safety evaluation for expressways: A comparative study for macroscopic and microscopic indicators. Traffic Inj. Prev. 2014;15(1):89-93. DOI:10.1080/15389588.2013.782400.

[10] Rajaram V, Subramanian SC. Heavy vehicle collision avoidance control in heterogeneous traffic using varying time headway. Mechatronics. 2018;50:328-340. DOI:10.1016/j.mechatronics.2017.11.010.

[11] Ward JR, et al. Extending time to collision for probabilistic reasoning in general traffic scenarios. Transp. Res. Part C Emerg. Technol. 2015;51:66-82. DOI:10.1016/j.trc.2014.11.002.

[12] Wang C, Quddus MA, Ison SG. The effect of traffic and road characteristics on road safety: A review and future research direction. Saf. Sci. 2013;57:264-275. DOI:10.1016/j.ssci.2013.02.012.

[13] Sohail A, et al. Data-driven approaches for road safety: A comprehensive systematic literature review. Saf. Sci. 2023;158:105949. DOI:10.1016/j.ssci.2022.105949.

[14] Lukusa MT, Hing Phoa FK. A horvitz-type estimation on incomplete traffic accident data analyzed via a zero-inflated poisson model. Accid. Anal. Prev. 2020;134:105235. DOI:10.1016/j.aap.2019.07.011.

[15] Luo Q, Liu C. Exploration of road closure time characteristics of tunnel traffic accidents: A case study in pennsylvania, usa. Tunn. Undergr. Space Technol. 2023;132:104894. DOI:10.1016/j.tust.2022.104894.

[16] D'Andrea E, et al. Real-time detection of traffic from twitter stream analysis. IEEE Trans. Intell. Transp. Syst. 2015;16(4):2269-2283. DOI:10.1109/TITS.2015.2404431.

[17] Gu Y, Qian ZS, Chen F. From twitter to detector: Real-time traffic incident detection using social media data. Transp. Res. Pt. C-Emerg. Technol. 2016;67:321-342. DOI:10.1016/j.trc.2016.02.011.

[18] Zheng M, et al. Network space analysis ‑ based identification of road traffic accident hotspots: A case study. Int. J. Crashworthiness. 2023;28(1):108-115. DOI:10.1080/13588265.2022.2109446.

[19] Jiang F, Yuen KKR, Lee EWM. A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions. Accid. Anal. Prev. 2020;141:105520. DOI:10.1016/j.aap.2020.105520.

[20] Wen X, et al. Applications of machine learning methods in traffic crash severity modelling: Current status and future directions. Transp. Rev. 2021;41(6):855-879. DOI:10.1080/01441647.2021.1954108.

[21] Wang L, Abdel-Aty M, Lee J. Safety analytics for integrating crash frequency and real-time risk modeling for expressways. Accid. Anal. Prev. 2017;104:58-64. DOI:10.1016/j.aap.2017.04.009.

[22] Yuan J, Abdel-Aty M. Approach-level real-time crash risk analysis for signalized intersections. Accid. Anal. Prev. 2018;119:274-289. DOI:10.1016/j.aap.2018.07.031.

[23] Zhang Z, et al. A deep learning approach for detecting traffic accidents from social media data. Transp. Res. Pt. C-Emerg. Technol. 2018;86:580-596. DOI:10.1016/j.trc.2017.11.027.

[24] Bao J, Liu P, Ukkusuri SV. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. Accid. Anal. Prev. 2019;122:239-254. DOI:10.1016/j.aap.2018.10.015.

[25] Goh K, et al. Factors affecting the probability of bus drivers being at-fault in bus-involved accidents. Accid. Anal. Prev. 2014;66:20-26. DOI:10.1016/j.aap.2013.12.022.

[26] Young SD, Wang W, Chakravarthy B. Crowdsourced traffic data as an emerging tool to monitor car crashes. JAMA Surg. 2019;154(8):777-778. DOI:10.1001/jamasurg.2019.1167.

[27] Chen J, Tao W. Traffic accident duration prediction using text mining and ensemble learning on expressways. Sci. Rep. 2022;12(1):21478. DOI:10.1038/s41598-022-25988-4.

[28] Gutierrez-Osorio C, Pedraza C. Modern data sources and techniques for analysis and forecast of road accidents: a review. Journal of Traffic and Transportation Engineering (English Edition). 2020;7(4):432-446. DOI:10.1016/j.jtte.2020.05.002.

[29] Deng M, et al. An analysis of physiological responses as indicators of driver takeover readiness in conditionally automated driving. Accid. Anal. Prev. 2024;195:107372. DOI:10.1016/j.aap.2023.107372.

[30] Moriya K, Matsushima S, Yamanishi K. Traffic risk mining from heterogeneous road statistics. IEEE Trans. Intell. Transp. Syst. 2018;19(11):3662-3675. DOI:10.1109/TITS.2018.2856533.

[31] Basso F, et al. A deep learning approach for real-time crash prediction using vehicle-by-vehicle data. Accid. Anal. Prev. 2021;162:106409. DOI:10.1016/j.aap.2021.106409.

[32] Yang Z, et al. Driving risk assessment using cluster analysis based on naturalistic driving data. 2014 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). 2014. p. 2584-2589. DOI:10.1109/ITSC.2014.6958104.

[33] Cao Q, et al. Reallocation of heterogeneous sensors on road networks for traffic accident detection. IEEE Trans. Instrum. Meas. 2023;72:1-11. DOI:10.1109/TIM.2023.3291790.

[34] Chang HL, et al. Tracking traffic congestion and accidents using social media data: A case study of Shanghai. Accid. Anal. Prev. 2022;169:106618. DOI:10.1016/j.aap.2022.106618.

[35] Lu H, et al. Using adverse weather data in social media to assist with city-level traffic situation awareness and alerting. Appl. Sci. 2018;8(7):1193. DOI:10.3390/app8071193.

[36] Sinnott RO, Yin S. Accident black spot identification and verification through social media. 2015 IEEE International Conference on Data Science and Data Intensive Systems (DSDIS). 2015. p. 17-24. DOI:10.1109/DSDIS.2015.34.

[37] Dabiri S, Heaslip K. Developing a twitter-based traffic event detection model using deep learning architectures. Expert Syst. Appl. 2019;118:425-439. DOI:10.1016/j.eswa.2018.10.017.

[38] Durduran SS. A decision making system to automatic recognize of traffic accidents on the basis of a gis platform. Expert Syst. Appl. 2010;37(12):7729-7736. DOI:10.1016/j.eswa.2010.04.068.

[39] Shafabakhsh GA, Famili A, Bahadori MS. Gis-based spatial analysis of urban traffic accidents: Case study in Mashhad, Iran. J. Traffic Transp. Eng.-Engl. Ed. 2017;4(3):290-299. DOI:10.1016/j.jtte.2017.05.005.

[40] Guo J, et al. Gps-based citywide traffic congestion forecasting using cnn-rnn and c3d hybrid model. Transportmetrica A. 2021;17(2):190-211. DOI:10.1080/23249935.2020.1745927.

[41] Shahzad M. Review of road accident analysis using gis technique. Int. J. Inj. Control Saf. Promot. 2020;27(4):472-481. DOI:10.1080/17457300.2020.1811732.

[42] Ding T Q, et al. Accident probability prediction and analysis of bus drivers based on occupational characteristics. Appl. Sci.-Basel. 2024;14(1):279. DOI:10.3390/app14010279.

[43] Zhang ZB, et al. An assessment of the relationship between driving skills and driving behaviors among chinese bus drivers. Adv. Mech. Eng. 2019;11(1). DOI:10.1177/1687814018824916.

[44] Nasri M, Aghabayk K. Assessing risk factors associated with urban transit bus involved accident severity: A case study of a middle east country. Int. J. Crashworthiness. 2021;26(4):413-423. DOI:10.1080/13588265.2020.1718465.

[45] Luan S, et al. Effects of built environment on bicycle wrong way riding behavior: A data -driven approach. Accid. Anal. Prev. 2020;144:105613. DOI:10.1016/j.aap.2020.105613.

[46] Liang XY, Meng XH, Zheng L. Investigating conflict behaviours and characteristics in shared space for pedestrians, conventional bicycles and e-bikes. Accid. Anal. Prev. 2021;158:106167. DOI:10.1016/j.aap.2021.106167.

[47] Amin-Naseri M, et al. Evaluating the reliability, coverage, and added value of crowdsourced traffic incident reports from waze. Transp. Res. Rec. 2018;2672(43):34-43. DOI:10.1177/0361198118790619.

[48] Li JB, et al. Feedback on a shared big dataset for intelligent tbm part i: Feature extraction and machine learning methods. Undergr. Space. 2023;11:1-25. DOI:10.1016/j.undsp.2023.01.001.

[49] Tao W, et al. Feature optimization method for white feather broiler health monitoring technology. Eng. Appl. Artif. Intell. 2023;123:106372. DOI:10.1016/j.engappai.2023.106372.

[50] Zhang ZS, et al. A study on the method for cleaning and repairing the probe vehicle data. IEEE Trans. Intell. Transp. Syst. 2013;14(1):419-427. DOI:10.1109/TITS.2012.2217378.

[51] Yan M, Shen Y. Traffic accident severity prediction based on random forest. Sustainability. 2022;14(3):1729. DOI:10.3390/su14031729.

[52] Madley-Dowd P, et al. The proportion of missing data should not be used to guide decisions on multiple imputation. J. Clin. Epidemiol. 2019;110:63-73. DOI:10.1016/j.jclinepi.2019.02.016.

[53] Dal Bianco G, et al. A practical and effective sampling selection strategy for large scale deduplication. IEEE Trans. Knowl. Data Eng. 2015;27(9):2305-2319. DOI:10.1109/TKDE.2015.2416734.

[54] Xie S. Feature extraction of auto insurance size of loss data using functional principal component analysis. Expert Syst. Appl. 2022;198:116780. DOI:10.1016/j.eswa.2022.116780.

[55] Xu J. A weighted linear discriminant analysis framework for multi-label feature extraction. Neurocomputing. 2018;275:107-120. DOI:10.1016/j.neucom.2017.05.008.

[56] Akaho S. Conditionally independent component analysis for supervised feature extraction. Neurocomputing. 2002;49(1):139-150. DOI:10.1016/S0925-2312(02)00518-0.

[57] Fattahi M, Moattar MH, Forghani Y. Locally alignment based manifold learning for simultaneous feature selection and extraction in classification problems. Knowledge-Based Syst. 2023;259:110088. DOI:10.1016/j.knosys.2022.110088.

[58] Sharma AK, et al. Hog transformation based feature extraction framework in modified resnet50 model for brain tumor detection. Biomed. Signal Process. Control. 2023;84:104737. DOI:10.1016/j.bspc.2023.104737.

[59] Suarez-Alvarez MM, et al. Statistical approach to normalization of feature vectors and clustering of mixed datasets. Proc. R. Soc. A-Math. Phys. Eng. Sci. 2012;468(2145):2630-2651. DOI:10.1098/rspa.2011.0704.

[60] Sengupta A, et al. A bayesian approach to quantifying uncertainties and improving generalizability in traffic prediction models. Transp. Res. Pt. C-Emerg. Technol. 2024;162:104585. DOI:10.1016/j.trc.2024.104585.

[61] Munkhdalai L, et al. Mixture of activation functions with extended min-max normalization for forex market prediction. IEEE Access. 2019;7:183680-183691. DOI:10.1109/ACCESS.2019.2959789.

[62] Jain S, Shukla S, Wadhvani R. Dynamic selection of normalization techniques using data complexity measures. Expert Syst. Appl. 2018;106:252-262. DOI:10.1016/j.eswa.2018.04.008.

[63] AlMamlook RE, et al. Comparison of machine learning algorithms for predicting traffic accident severity. 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT). 2019. p. 272-276. DOI:10.1109/JEEIT.2019.8717393.

[64] Parsa AB, et al. Real-time accident detection: coping with imbalanced data. Accid. Anal. Prev. 2019;129:202-210. DOI:10.1016/j.aap.2019.05.014.

[65] Shangguan AQ, et al. Traffic accident severity prediction based on oversampling and cnn for imbalanced data. 2021 Proceedings of The 40th Chinese Control Conference(CCC). 2021. p. 7004-7008. DOI:10.23919/CCC52363.2021.9549759.

[66] Wang C, et al. An analysis of fatal accident rates of passenger cars on urban roads considering imbalanced data samples. Journal of Transport Information and Safety. 2023;41(5):43-53. DOI:10.3963/j.jssn.1674-4861.2023.05.005.

[67] Ferrer CA, Aragón E. Note on "a comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance". Inf. Sci. 2023;630:322-324. DOI:10.1016/j.ins.2022.10.005.

[68] Shi X, et al. Solving the data imbalance problem of p300 detection via random under-sampling bagging svms. 2015 International Joint Conference on Neural Networks (IJCNN). 2015. p. 1-5. DOI:10.1109/IJCNN.2015.7280834.

[69] Pereira RM, Costa YMG, Silla Jr. CN. Mltl: a multi-label approach for the tomek link undersampling algorithm. Neurocomputing. 2020;383:95-105. DOI:10.1016/j.neucom.2019.11.076.

[70] Mujalli RO, López G, Garach L. Bayes classifiers for imbalanced traffic accidents datasets. Accid. Anal. Prev. 2016;88:37-51. DOI:10.1016/j.aap.2015.12.003.

[71] Tahir M, et al. Real-time event-driven road traffic monitoring system using cctv video analytics. IEEE Access. 2023;11:139097-139111. DOI:10.1109/ACCESS.2023.3340144.

[72] Zhao C, et al. Unsupervised anomaly detection based method of risk evaluation for road traffic accident. Appl. Intell. 2023;53(1):369-384. DOI:10.1007/s10489-022-03501-8.

[73] Monisha M, et al. A new generalization of the zero-truncated negative binomial distribution by a lagrange expansion with associated regression model and applications. Int. J. Data Sci. Anal. 2023:1-5. DOI:10.1007/s41060-023-00449-x.

[74] Ding Y, et al. Deep imbalanced regression using cost-sensitive learning and deep feature transfer for bearing remaining useful life estimation. Appl. Soft Comput. 2022;127:109271. DOI:10.1016/j.asoc.2022.109271.

[75] Lv WF, Cui W, Huang J. Research on a datex iii based dynamic traffic information publish platform. 2008 2nd International Symposium on Intelligent Information Technology Application. 2008. p. 412-416. DOI:10.1109/IITA.2008.81.

[76] Jakovljevic D, Balen J, Vidovic K. Integration of traffic and travel data exchange in command and control platform. 2016 1st International Conference on Smart Systems and Technologies (SST). 2016. p. 281-286. DOI:10.1109/SST.2016.7765674.

[77] Anonymous. Traffic management data dictionary (tmdd) and message set for external traffic management center communications (ms/etmcc) guide. ITE J. 2001;71(4):24-24.

[78] Mohammed SJ, Hasson ST, IEEE. Modeling and simulation of data dissemination in vanet based on a clustering approach. 2022 International Conference on Computer Science and Software Engineering (CSASE). 2022. p. 54-59. DOI:10.1109/CSASE51777.2022.9759671.

[79] Raut YC, et al. Early alert system using relative positioning in vehicular ad-hoc network. 2014 Annual International Conference on Emerging Research Areas - Magnetics, Machines and Drives (AICERA/ICMMD). 2014. p. 1-8. DOI:10.1109/AICERA.2014.6908167.

[80] Sou SI, Tonguz OK. Enhancing vanet connectivity through roadside units on highways. IEEE Trans. Veh. Technol. 2011;60(8):3586-3602. DOI:10.1109/TVT.2011.2165739.

[81] Nasr E, et al. An iot approach to vehicle accident detection, reporting, and navigation. 2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET). 2016. p. 231-236. DOI:10.1109/IMCET.2016.7777457.

[82] Kumar A, Das D, IEEE. Intelligentchain: blockchain and machine learning based intelligent security application for internet of vehicles (iov). 2022 IEEE 95th Vehicular Technology Conference: (VTC-Spring). 2022. p. 1-5. DOI:10.1109/VTC2022-Spring54318.2022.9860946.

[83] Marcillo P, Caraguay A, Hernandez-Alvarez M. A systematic literature review of learning-based traffic accident prediction models based on heterogeneous sources. Appl. Sci.-Basel. 2022;12(9):4529. DOI:10.3390/app12094529.

[84] Wang L, et al. Real-time crash prediction for expressway weaving segments. Transp. Res. Pt. C-Emerg. Technol. 2015;61:1-10. DOI:10.1016/j.trc.2015.10.008.

[85] Harb R, et al. Exploring precrash maneuvers using classification trees and random forests. Accid. Anal. Prev. 2009;41(1):98-107. DOI:10.1016/j.aap.2008.09.009.

[86] Li P, Abdel-Aty M, Yuan J. Real-time crash risk prediction on arterials based on lstm-cnn. Accid. Anal. Prev. 2020;135:105371. DOI:10.1016/j.aap.2019.105371.

[87] Greibe P. Accident prediction models for urban roads. Accid. Anal. Prev. 2003;35(2):273-285. DOI:10.1016/S0001-4575(02)00005-2.

[88] Zhang J, et al. Prediction of urban expressway total traffic accident duration based on multiple linear regression and artificial neural network. 2019 5th International Conference on Transportation Information and Safety (ICTIS). 2019. p. 503-510. DOI:10.1109/ictis.2019.8883690.

[89] Getahun KA. Time series modeling of road traffic accidents in amhara region. J. Big Data. 2021;8(1):102. DOI:10.1186/s40537-021-00493-z.

[90] Chen ZY, et al. Macro-level accident fatality prediction using a combined model based on arima and multivariable linear regression. 2016 4th IEEE International Conference on Progress in Informatics and Computing (IEEE PIC). 2016. p. 133-137. DOI:10.1109/PIC.2016.7949481.

[91] Donnelly-Swift E, Kelly A. Factors associated with single-vehicle and multi-vehicle road traffic collision injuries in ireland. Int. J. Inj. Control Saf. Promot. 2016;23(4):351-361. DOI:10.1080/17457300.2015.1047861.

[92] Wong CK. Designs for safer signal-controlled intersections by statistical analysis of accident data at accident blacksites. IEEE Access. 2019;7:111302-111314. DOI:10.1109/ACCESS.2019.2928038.

[93] Awad HA, Parry T. Investigation the effect of pavement condition characteristics on bend segments accident frequency: application of fixed and random parameters negative binomial models. 2018 International conference on transportation and development. 2018. p. 165-176. DOI:10.1061/9780784481554.017.

[94] Chin HC, Quddus MA. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accid. Anal. Prev. 2003;35(2):253-259. DOI:10.1016/S0001-4575(02)00003-9.

[95] Karacasu M, Ergul B, Yavuz AA. Estimating the causes of traffic accidents using logistic regression and discriminant analysis. Int. J. Inj. Control Saf. Promot. 2014;21(4):305-312. DOI:10.1080/17457300.2013.815632.

[96] Gu X, et al. Utilizing uav video data for in-depth analysis of drivers' crash risk at interchange merging areas. Accid. Anal. Prev. 2019;123:159-169. DOI:10.1016/j.aap.2018.11.010.

[97] Çelik AK, Oktay E. A multinomial logit analysis of risk factors influencing road traffic injury severities in the erzurum and kars provinces of turkey. Accid. Anal. Prev. 2014;72:66-77. DOI:10.1016/j.aap.2014.06.010.

[98] Asare IO, Mensah AC. Crash severity modelling using ordinal logistic regression approach. Int. J. Inj. Control Saf. Promot. 2020;27(4):412-419. DOI:10.1080/17457300.2020.1790615.

[99] Fountas G, Anastasopoulos PC, Abdel-Aty M. Analysis of accident injury-severities using a correlated random parameters ordered probit approach with time variant covariates. Anal. Methods Accid. Res. 2018;18:57-68. DOI:10.1016/j.amar.2018.04.003.

[100] Hyodo S, Hasegawa K. Factors affecting analysis of the severity of accidents in cold and snowy areas using the ordered probit model. Asian Transp. Stud. 2021;7:100035. DOI:10.1016/j.eastsj.2021.100035.

[101] Abdel-Aty M. Analysis of driver injury severity levels at multiple locations using ordered probit models. J. Saf. Res. 2003;34(5):597-603. DOI:10.1016/j.jsr.2003.05.009.

[102] Lee J, et al. Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: seoul city study. Appl. Sci.-Basel. 2020;10(1):129. DOI:10.3390/app10010129.

[103] Chang LY, Chen WC. Data mining of tree-based models to analyze freeway accident frequency. J. Saf. Res. 2005;36(4):365-375. DOI:10.1016/j.jsr.2005.06.013.

[104] Zhao GM, Zhou WH, Li J. Regression analysis of association between vehicle performance and driver casualty risk in traffic accidents. 2015 International Conference on Transportation Information and Safety (ICTIS). 2015. p. 345-349. DOI:10.1109/ICTIS.2015.7232070.

[105] Chang LY, Wang HW. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid. Anal. Prev. 2006;38(5):1019-1027. DOI:10.1016/j.aap.2006.04.009.

[106] Balawi M, Tenekeci G. Time series traffic collision analysis of london hotspots: patterns, predictions and prevention strategies. Heliyon. 2024;10(4):e25710. DOI:10.1016/j.heliyon.2024.e25710.

[107] de Grange L, et al. Estimating the impact of incidents on urban controlled-access highways: An empirical analysis. Appl. Econ. 2017;49(18):1763-1773. DOI:10.1080/00036846.2016.1226489.

[108] Dereli MA, Erdogan S. A new model for determining the traffic accident black spots using gis-aided spatial statistical methods. Transp. Res. Pt. A-Policy Pract. 2017;103:106-117. DOI:10.1016/j.tra.2017.05.031.

[109] Yu L, et al. Deep spatio-temporal graph convolutional network for traffic accident prediction. Neurocomputing. 2021;423:135-147. DOI:10.1016/j.neucom.2020.09.043.

[110] Abellán J, López G, de Oña J. Analysis of traffic accident severity using decision rules via decision trees. Expert Syst. Appl. 2013;40(15):6047-6054. DOI:https://doi.org/10.1016/j.eswa.2013.05.027.

[111] Azhar A, et al. Classification of driver injury severity for accidents involving heavy vehicles with decision tree and random forest. Sustainability. 2022;14(7):4101. DOI:10.3390/su14074101.

[112] Zhang X, Fan L. A decision tree approach for traffic accident analysis of saskatchewan highways. 2013 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE). 2013. p. 1-4. DOI:10.1109/CCECE.2013.6567833.

[113] Wu W, et al. Economic development, demographic characteristics, road network and traffic accidents in zhongshan, china: gradient boosting decision tree model. Transportmetrica A. 2020;16(3):359-387. DOI:10.1080/23249935.2020.1711543.

[114] Zhang WH, Liu T, Yi J. Exploring the spatiotemporal characteristics and causes of rear-end collisions on urban roadways. Sustainability. 2022;14(18):11761. DOI:10.3390/su141811761.

[115] Xiong X, Chen L, Liang J. A new framework of vehicle collision prediction by combining svm and hmm. IEEE Trans. Intell. Transp. Syst. 2018;19(3):699-710. DOI:10.1109/TITS.2017.2699191.

[116] Lv YS, et al. Real-time highway accident prediction based on support vector machines. 2009 21st Chinese Control and Decision Conference. 2009. p. 4403-4407. DOI:10.1109/CCDC.2009.5192409.

[117] Cui KB, Du YS, Hu Z. Short-term load forecasting based on the bkf-svm. 2009 International Conference on Networks Security, Wireless Communications and Trusted Computing. 2009. p. 528-531. DOI:10.1109/NSWCTC.2009.170.

[118] Kim EJ, et al. Application of naive bayesian approach in detecting reproducible fatal collision locations on freeway. PLoS One. 2021;16(5):e0251866. DOI:10.1371/journal.pone.0251866.

[119] Kumar N, Acharya D, Lohani D. An iot-based vehicle accident detection and classification system using sensor fusion. IEEE Internet Things J. 2021;8(2):869-880. DOI:10.1109/JIOT.2020.3008896.

[120] Bahiru TK, et al. Comparative study on data mining classification algorithms for predicting road traffic accident severity. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). 2018. p. 1655-1660. DOI:10.1109/ICICCT.2018.8473265.

[121] Ghadi M, Torok A, Tanczos K. Integration of probability and clustering based approaches in the field of black spot identification. Period. Polytech.-Civ. Eng. 2019;63(1):46-52. DOI:10.3311/PPci.11753.

[122] Kumar S, Toshniwal D. A data mining approach to characterize road accident locations. J. Mod. Transp. 2016;24(1):62-72. DOI:10.1007/s40534-016-0095-5.

[123] Anderson TK. Kernel density estimation and k-means clustering to profile road accident hotspots. Accid. Anal. Prev. 2009;41(3):359-364. DOI:10.1016/j.aap.2008.12.014.

[124] Yaman TT, Bilgic E, Esen MF. Analysis of traffic accidents with fuzzy and crisp data mining techniques to identify factors affecting injury severity. J. Intell. Fuzzy Syst. 2022;42(1):575-592. DOI:10.3233/JIFS-219213.

[125] Zhang YF, et al. Exploring spatiotemporal patterns of expressway traffic accidents based on density clustering and bayesian network. ISPRS Int. J. Geo-Inf. 2023;12(2):73. DOI:10.3390/ijgi12020073.

[126] Zhang X, et al. Traffic accident location study based on ad-dbscan algorithm with adaptive parameters. 2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD). 2023. p. 1160-1165. DOI:10.1109/CSCWD57460.2023.10152613.

[127] Zhong WF, Du LJ. Predicting traffic casualties using support vector machines with heuristic algorithms: A study based on collision data of urban roads. Sustainability. 2023;15(4):2944. DOI:10.3390/su15042944.

[128] Montazeri GM, Fotouhi A. Traffic condition recognition using the k-means clustering method. Sci. Iran. 2011;18(4):930-937. DOI:10.1016/j.scient.2011.07.004.

[129] Syahputri K, et al. Clustering the vulnerability of traffic accidents in medan city with fuzzy c-means algorithm. 2nd Talenta Conference on Engineering, Science and Technology. 2020;801(1):012030. DOI:10.1088/1757-899X/801/1/012030.

[130] Nogueira P, et al. Learning from accidents: spatial intelligence applied to road accidents with insights from a case study in setubal district, portugal. ISPRS Int. J. Geo-Inf. 2023;12(3):93. DOI:10.3390/ijgi12030093.

[131] Khan SW, et al. Anomaly detection in traffic surveillance videos using deep learning. Sensors. 2022;22(17):6563. DOI:10.3390/s22176563.

[132] Zhao HT, et al. Deep learning-based prediction of traffic accidents risk for internet of vehicles. China Commun. 2022;19(2):214-224. DOI:10.23919/JCC.2022.02.017.

[133] Pérez-Sala L, et al. Deep learning model of convolutional neural networks powered by a genetic algorithm for prevention of traffic accidents severity. Chaos Solitons Fractals. 2023;169:113245. DOI:10.1016/j.chaos.2023.113245.

[134] Huang T, Wang S, Sharma A. Highway crash detection and risk estimation using deep learning. Accid. Anal. Prev. 2020;135:105392. DOI:10.1016/j.aap.2019.105392.

[135] Chang YL, et al. Consolidated convolutional neural network for hyperspectral image classification. Remote Sens. 2022;14(7):1571. DOI:10.3390/rs14071571.

[136] Sun YS, et al. Detecting anomalous traffic behaviors with seasonal deep kalman filter graph convolutional neural networks. J. King Saud Univ.-Comput. Inf. Sci. 2022;34(8):4729-4742. DOI:10.1016/j.jksuci.2022.05.017.

[137] Wang S, et al. Traffic accident risk prediction via multi-view multi-task spatio-temporal networks. IEEE Trans. Knowl. Data Eng. 2021;35(12):12323-12336. DOI:10.1109/TKDE.2021.3135621.

[138] Sameen MI, Pradhan B. Severity prediction of traffic accidents with recurrent neural networks. Appl. Sci.-Basel. 2017;7(6):476. DOI:10.3390/app7060476.

[139] Shang Q, Xie T, Yu Y. Prediction of duration of traffic incidents by hybrid deep learning based on multi-source incomplete data. Int. J. Environ. Res. Public Health. 2022;19(17):10903. DOI:10.3390/ijerph191710903.

[140] Ren H, et al. A deep learning approach to the citywide traffic accident risk prediction. 2018 21st IEEE International Conference on Intelligent Transportation Systems (ITSC). 2018. p. 3346-3351. DOI: 10.1109/ITSC.2018.8569437.

[141] Lin DJ, et al. Intelligent traffic accident prediction model for internet of vehicles with deep learning approach. IEEE Trans. Intell. Transp. Syst. 2022;23(3):2340-2349. DOI:10.1109/TITS.2021.3074987.

[142] Hu Z, Zhou J, Zhang E. Improving traffic safety through traffic accident risk assessment. Sustainability. 2023;15(4):3748. DOI:10.3390/su15043748.

[143] Senarath Y, et al. Practitioner-centric approach for early incident detection using crowdsourced data for emergency services. 2021 21st IEEE International Conference on Data Mining (IEEE ICDM). 2021. p. 1318-1323. DOI:10.1109/ICDM51629.2021.00164.

[144] Yang D, et al. Freeway accident detection and classification based on the multi-vehicle trajectory data and deep learning model. Transp. Res. Pt. C-Emerg. Technol. 2021;130:103303. DOI:10.1016/j.trc.2021.103303.

[145] Wu MY, et al. A multi-attention dynamic graph convolution network with cost-sensitive learning approach to road-level and minute-level traffic accident prediction. IET Intell. Transp. Syst. 2023;17(2):270-284. DOI:10.1049/itr2.12254.

[146] Kashifi MT, Al-Turki M, Sharify AW. Deep hybrid learning framework for spatiotemporal crash prediction using big traffic data. International Journal of Transportation Science and Technology (IJTST). 2022;12(3):793-808. DOI:10.1016/j.ijtst.2022.07.003.

[147] de Medrano R, Aznarte JL. A new spatio-temporal neural network approach for traffic accident forecasting. Appl. Artif. Intell. 2021;35(10):782-801. DOI:10.1080/08839514.2021.1935588.

[148] Lv YS, et al. Real-time highway traffic accident prediction based on the k-nearest neighbor method. 2009 International Conference on Measuring Technology and Mechatronics Automation. 2009. p. 547-550. DOI:10.1109/ICMTMA.2009.657.

[149] Lv YS, et al. Traffic flow prediction with big data: a deep learning approach. IEEE Trans. Intell. Transp. Syst. 2015;16(2):865-873. DOI:10.1109/TITS.2014.2345663.

[150] Ma XL, et al. Large-scale transportation network congestion evolution prediction using deep learning theory. PLoS One. 2015;10(3):e0119044. DOI:10.1371/journal.pone.0119044.

[151] Yang ZK, Zhang WP, Feng J. Predicting multiple types of traffic accident severity with explanations: A multi-task deep learning framework. Saf. Sci. 2022;146:105522. DOI:10.1016/j.ssci.2021.105522.

[152] Cho I, et al. Reinforcement learning for predicting traffic accidents. 2023 5th International Conference on Artificial Intelligence in Information and Communication (ICAIIC). 2023. p. 684-688. DOI:10.1109/ICAIIC57133.2023.10067034.

[153] Ji XY, et al. Digital twin empowered model free prediction of accident-induced congestion in urban road networks. 2022 IEEE 95th Vehicular Technology Conference: (VTC-Spring). 2022. p. 1-6. DOI:10.1109/VTC2022-Spring54318.2022.9860491.

[154] Zhang HT, et al. Overfitting and underfitting analysis for deep learning based end-to-end communication systems. 2019 11th IEEE International Conference on Wireless Communications and Signal Processing (WCSP). 2019. p. 1-6. DOI:10.1109/wcsp.2019.8927876.

[155] Watanabe S, Yamana H, IEEE. Overfitting measurement of deep neural networks using no data. 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA). 2021. p. 1-10. DOI:10.1109/DSAA53316.2021.9564119.

[156] Li QP, Yan M, Xu J. Optimizing convolutional neural network performance by mitigating underfitting and overfitting. 2021 IEEE/ACIS 19th International Conference on Computer and Information Science (ICIS). 2021. p. 126-131. DOI:10.1109/ICIS51600.2021.9516868.

[157] Liu Y, et al. Ensemble learning with correlation-based penalty. 2014 IEEE 12th International Conference on Dependable, Autonomic and Secure Computing (DASC). 2014. p. 350-353. DOI:10.1109/DASC.2014.69.

[158] Liu Y, IEEE. New discoveries in balanced ensemble learning. 2012 International Joint Conference on Neural Networks (IJCNN). 2012. p. 1-8. DOI: 10.1109/IJCNN.2012.6252423.

[159] Guo Y, et al. Research review of space-frequency domain image enhancement methods. Computer Engineering and Application. 2022;58(11):23-32. DOI:1002-8331(2022)58:11<23:KPYTXZ>2.0.TX;2-4.

[160] Wu XX, Liu JG. A new early stopping algorithm for improving neural network generalization. 2009 2nd International Conference on Intelligent Computation Technology and Automation. 2009. p. 15-18. DOI:10.1109/ICICTA.2009.11.

[161] Chen H, et al. Image recognition algorithm based on artificial intelligence. Neural Comput. Appl. 2022;34(9):6661-6672. DOI:10.1007/s00521-021-06058-8.

[162] Scala F, et al. A general approach to dropout in quantum neural networks. Adv. Quantum Technol. 2023:2300220. DOI:10.1002/qute.202300220.

[163] Spiliotis A, et al. Integration and field evaluation of an iov system for enhancing road safety. Appl. Sci.-Basel. 2022;12(23):12262. DOI:10.3390/app122312262.

[164] Essa M, Sayed T. Self-learning adaptive traffic signal control for real-time safety optimization. Accid. Anal. Prev. 2020;146:105713. DOI:10.1016/j.aap.2020.105713.

何美玲, 孟光荣, 武晓晖, 韩珣, 范江洋

基于多源数据的道路交通事故预测：系统综述

摘要：
随着城市化进程的加快和道路交通量的快速增长，准确的交通事故预测已成为提高道路安全和交通效率的关键。然而，交通事故预测是一个复杂而多方面的问题，需要综合考虑人、车辆、道路和环境等多种因素。本文基于多源数据对交通事故预测进行了详细分析，通过综合考虑数据来源、数据处理和预测方法，从多角度全面的介绍了交通事故预测的相关内容，帮助读者了解不同数据和方法之间的特点，了解事故预测的流程，理解其中的关键技术和方法。文章最后总结了目前交通事故预测研究中的主要挑战与展望，例如，不同部门和领域之间缺乏有效的数据共享，对多源数据的整合构成了重大挑战。未来，将深度学习模型与社交媒体和车辆网络数据等时间敏感数据相结合，可以有效提高实时事故预测的准确性。

关键词：
多源数据；交通事故；数据处理；统计学习；机器学习；深度学习