Heliyon 10 (2024) e35595

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Prediction and interpretive of motor vehicle traffic crashes severity based on random forest optimized by meta-heuristic algorithm

Xing Wang [a], Yikun Su [a,*], Zhizhe Zheng [a], Liang Xu [b]

[a] School of Civil Engineering and Transportation, Northeast Forestry University, Harbin, 150040, China
[b] School of Civil Engineering, Changchun Institute of Technology, Changchun, 130012, China

ARTICLE INFO

Keywords:
Vehicle crashes
Crash severity classification
Random forest
Meta-heuristic algorithms
GIS
Explainable machine learning

ABSTRACT

Providing accurate prediction of the severity of traffic collisions is vital to improve the efficiency of emergencies and reduce casualties, accordingly improving traffic safety and reducing traffic congestion. However, the issue of both the predictive accuracy of the model and the interpretability of predicted outcomes has remained a persistent challenge. We propose a Random Forest optimized by a Meta-heuristic algorithm prediction framework that integrates the spatiotemporal characteristics of crashes. Through predictive analysis of motor vehicle traffic crash data on interstate highways within the United States in 2020, we compared the accuracy of various ensemble models and single-classification prediction models. The results show that the Random Forest (RF) model optimized by the Crown Porcupine Optimizer (CPO) has the best prediction results, and the accuracy, recall, f1 score, and precision can reach more than 90 %. We found that factors such as Temperature and Weather are closely related to vehicle traffic crashes. Closely related indicators were analyzed interpretatively using a geographic information system (GIS) based on the characteristic importance ranking of the results. The framework enables more accurate prediction of motor vehicle traffic crashes and discovers the important factors leading to motor vehicle traffic crashes with an explanation. The study proposes that in some areas consideration should be given to adding measures such as nighttime lighting devices and nighttime fatigue driving alert devices to ensure safe driving. It offers references for policymakers to address traffic management and urban development issues.

1. Introduction

According to data from the National Highway Traffic Safety Administration (NHTSA), the number of traffic accident fatalities in the United States surged by 7.2 % in 2020 compared to the previous year, marking a 13-year high, despite a relatively lower frequency of driving occurrences [1]. In aggregate, 38,680 dead succumbed to traffic crashes in the United States in 2020, representing the highest figure since 2007. However, the situation is even graver than it sounds, as the mileage driven in the United States decreased by 13 % from the previous year. Motor vehicle traffic crashes not only affect the quality of traffic and cause congestion but also lead to substantial property damage and casualties, imposing a significant burden on families and society [2,3]. By accurately predicting the

[*] Corresponding author.
E-mail address: suyk@nefu.edu.cn (Y. Su).

https://doi.org/10.1016/j.heliyon.2024.e35595
Received 14 May 2024; Received in revised form 24 July 2024; Accepted 31 July 2024
Available online 8 August 2024

severity of crashes and analyzing the impact of different characteristic factors on crashes, the losses associated with motor vehicle traffic crash severity can be effectively mitigated.

Due to the significance of predicting and identifying key factors of motor vehicle traffic crash severity, there has been a quantity of research investigating the relationship between crash severity outcomes and their related risk factors such as the traffic volume, road design, and environmental characteristics, etc [4–6]. Current studies mainly analyze the severity of vehicle collisions from provincial or regional perspectives, and the effects of multiple factors on the severity of vehicle traffic crashes have been analyzed through studies of regions such as central Taiwan [7], Miami-Dade County [8], California [9], and seven highways in Washington State [10]. However, a comprehensive nationwide spatial distribution analysis of vehicle collision severity is lacking. Moreover, substantial variations exist in the correlation between collision severity and influencing factors across different regions. The driving environment, time of travel, driver characteristics, road design, and vehicle type can all potentially be important factors affecting the severity of collisions [11–14]. Traffic flows are usually more intense in areas with higher population densities and higher levels of Gross Domestic Product (GDP) development, which will increase the probability of crashes. Thus, regional population density and level of GDP development may also become important factors affecting the severity of vehicle traffic crashes. However, existing research rarely includes social factors such as population density and GDP development level as predictive indicators. The accuracy of motor vehicle traffic crash severity prediction models has attracted much attention. With the development of computer science, data mining, and machine learning algorithms have emerged, providing an innovative and effective approach for studying the motor vehicle traffic crash severity. This has significantly improved the accuracy of predicting the motor vehicle traffic crash severity [15–17]. Machine learning models can extract important knowledge from vast amounts of intricate and diverse data, and they have attracted much attention in the field of traffic safety due to their outstanding predictive capabilities [18] Among the multitude of machine learning models that are widely employed, there are Support Vector Machines (SVM) [19], Back Propagation Neural Networks (BP) [20], Decision Trees (DT) [21], Bayesian Networks (BN), Extreme Gradient Boosting (XGBoos) [22], and Random Forests (RF) [23], etc. In recent years, deep learning frameworks have emerged based on machine learning to predict motor vehicle traffic crash severity [24,25]. Deep learning is known for effectively addressing pattern recognition issues in unstructured data. However, for structured data such as tabular data, traditional machine learning models are widely used. Traditional machine learning algorithms are typically more interpretable, allowing us to comprehend the inherent relationships that exist in the prediction process. Ziakopoulos et al. [26–29]effectively identified the occurrence of accidents using the XGBoost algorithm and elucidated the influence of individual characteristics on accident occurrence with the Shapley. Upon this, a network analysis framework based on XGBoost was developed to identify the critical factors influencing the severity of accidents. Umer [30] explored the problem of predicting the severity of motor vehicle traffic crashes by comparing an integrated tree-based learning model with a traditional statistical model. He concluded that the random forest model exhibited optimal performance in terms of classification accuracy. Ramya [31] used the Random Forest algorithm to assess the performance of Artificial Neural Networks (ANN) and Decision Trees in predicting the motor vehicle traffic crash severity. It is worth noting that Random Forest achieved the best performance in most of the studies, and machine learning models with high predictive accuracy were commonly chosen to calculate feature importance.

The probability of motor vehicle traffic crashes varies spatially and temporally [32,33]. Thus, the temporal and spatial characteristics of motor vehicle traffic crashes must be considered when studying the impact of road traffic safety and transport system properties on a specific area and their interrelationships with neighboring areas. GIS technology is favored by scholars for its spatial analysis capabilities and powerful visualization functions [34]. GIS enables the visualization of motor vehicle traffic crash distributions, offering insights into spatial patterns [35,36]. Additionally, by leveraging multiple spatial analysis tools within GIS and integrating data from various sources a comprehensive exploration of the spatial distribution of motor vehicle traffic crashes can be conducted [37,38]. This approach allows for a detailed analysis of temporal and spatial characteristics of motor vehicle traffic crash severity.

In summary, based on previous research, this paper aims to develop a framework for predicting the motor vehicle traffic crash severity using a Random Forest model combined with Meta-Heuristic Algorithms. The integration of Geographic Information Systems (GIS) and machine learning techniques helps address the limitations observed in previous models for predicting motor vehicle traffic crash crash severity. This integration enhances the accuracy of prediction models and facilitates a comprehensive analysis of the interplay between various influencing factors and the incidence of crashes. We used crash data from 2020 for interstate highways in the Contiguous United States (All contiguous areas except Alaska or Hawaii) for our analysis. Raw statistics contain address, time, weather, and road attributes in text form and need to be filtered for valid information in the raw data. We used geographic coordinate information to pinpoint each crash on the GIS and integrated social factors that influence crashes, such as population density and GDP development level. The data were processed by the Synthetic Minority Oversampling Technique (SMOTE) to address the structural imbalance of the data. A categorical prediction model using a meta-heuristic algorithm to optimize random forests was employed to analyze the prediction of motor vehicle traffic crash severity. The prediction results were then compared with those of a single categorical prediction model to determine the model with the highest prediction accuracy. According to the ranking of the importance of indicators in predicting results, along with the visualization function of GIS, the significant indicators influencing motor vehicle traffic crashes are analyzed in an explanatory manner. The study results can be used to predict motor vehicle traffic crashes more accurately and discover the important factors leading to motor vehicle traffic crashes with an explanation. It offers references for policymakers to address traffic management and urban development issues.

The rest of the paper is organized as follows. In Section 2, we describe the sample sources and the identification process. The method's structure and computational procedures are described in Section 3. Model application and experimental results are provided in Section 4. Section 5 includes conclusions.

2. Data description and preprocessing

2.1. Data description

For this study, we utilized a dataset of crashes that occurred on interstate highways in the Contiguous United States in 2020. The dataset was obtained from Kaggle, a well-known machine-learning competition platform [39]. This dataset was collected in real-time using multiple traffic APIs that cover 49 states of the USA. The dataset contains 46 features of information describing each crash. One of these features is the crash severity classification (class), while the others can be divided into five sections: basic information about the crash (ID, Source, Description, etc.), crash location (Start-Lat, Start-Log, City, etc.), time of crash (Start-Time, End-Time, Sunrise--Sunset, etc.), environment (Temperature, Humidity, Weather, etc.), and crash roadway (Junction, Traffic-Calming, Traffic-Signal, etc.).

2.2. Data cleaning and transformation

Preprocessing motor vehicle traffic crash severity data plays a vital role in the machine learning process. The dataset contains multiple descriptions of crashes, but not every line of data can be used as an indicator for analysis, such as "Description", "City", "County Description", "City", "County", etc. Outliers, redundancies, and missing values must be removed from the data before analyzing it to ensure that the model receives accurate input data and makes precise predictions. We extracted 110,778 interstate highway collision records from all vehicle crash data in 2020 and conducted data structure analysis to eliminate 136 outliers and 6131 missing values from the original dataset. we finally obtained 104,511 data points, including environmental factors, road factors, time factors, social factors, and location factors. Table 1 displays all indicators post-screening and includes informative descriptions of each indicator.

Digitization and normalization of input data are necessary because the training and analysis of machine learning models require data with well-defined and consistent rules. Therefore, we use numerical values instead of categorical variables. We categorized the weather into seven categories based on its effect on regular vehicle travel. Road conditions were categorized as either present (1) or absent (0). Time information includes whether it is a rest day, season, and time of day [40]. In addition to this, the dataset includes

Table 1
Variable summary.

Category	NO.	Features	Definitions/Details
Environmental factor	1	Temperature (F)	Shows the temperature (in Fahrenheit)
	2	Wind-Chill (F)	Shows the wind chill (in Fahrenheit).
	3	Humidity (%)	Shows the humidity (in percentage).
	4	Pressure (in)	Shows the air pressure (in inches).
	5	Visibility (mi)	Shows visibility (in miles).
	6	Wind-Speed (mph)	Shows wind speed (in miles per hour).
	7	Precipitation (in)	Shows precipitation amount in inches
	8	Weather	Shows the weather conditions (rain, snow, thunderstorm, fog, etc.) 1 = (Fair, Windy, etc.) 2 = (Mostly Cloudy, Partly Cloudy, etc.) 3 = (Thunder, Rain Shower, etc.) 4 = (Rain, Precipitation, etc.) 5 = (Snow, Wintry Mix, etc.) 6 = (Mist, Dust Whirlwinds, etc.) 7 = (Heavy T-Storm, Heavy Drizzle, etc.)
Road factor	9	Junction	A POI annotation indicates the presence of a junction in a nearby location.
	10	Stop	A POI annotation indicates the presence of a stop in a nearby location.
	11	Traffic-Calming	A POI annotation indicates the presence of traffic calming in a nearby location.
	12	Traffic-Signal	A POI annotation indicates the presence of a traffic signal in a nearby location.
Temporal factor	13	Week	1 = Weekend (Saturday, Sunday) 0 = Weekday (Monday, Tuesday, Wednesday, Thursday, Friday)
	14	Season	1 = spring (March, April, May) 2 = summer (June, July, August) 3 = autumn (September, October, November) 4 = winter(December, January, February)
	15	Time of Day	1 = 6–9 am 2 = 4–7 pm 3 = 8 p.m.–5 am 4 = other
Social factor	16	Population Density	The ratio of total state population to area
	17	GDP	GDP of total state
Location Factor	18	Crash-Location	The crash location is reflected in the GIS
Severity	19	Crash severity	1 = Minor impact 2 = Moderate impact 3 = Impactful 4 = Significant impact

Please refer to Refs. [41,42] for POIs and other details in the table.

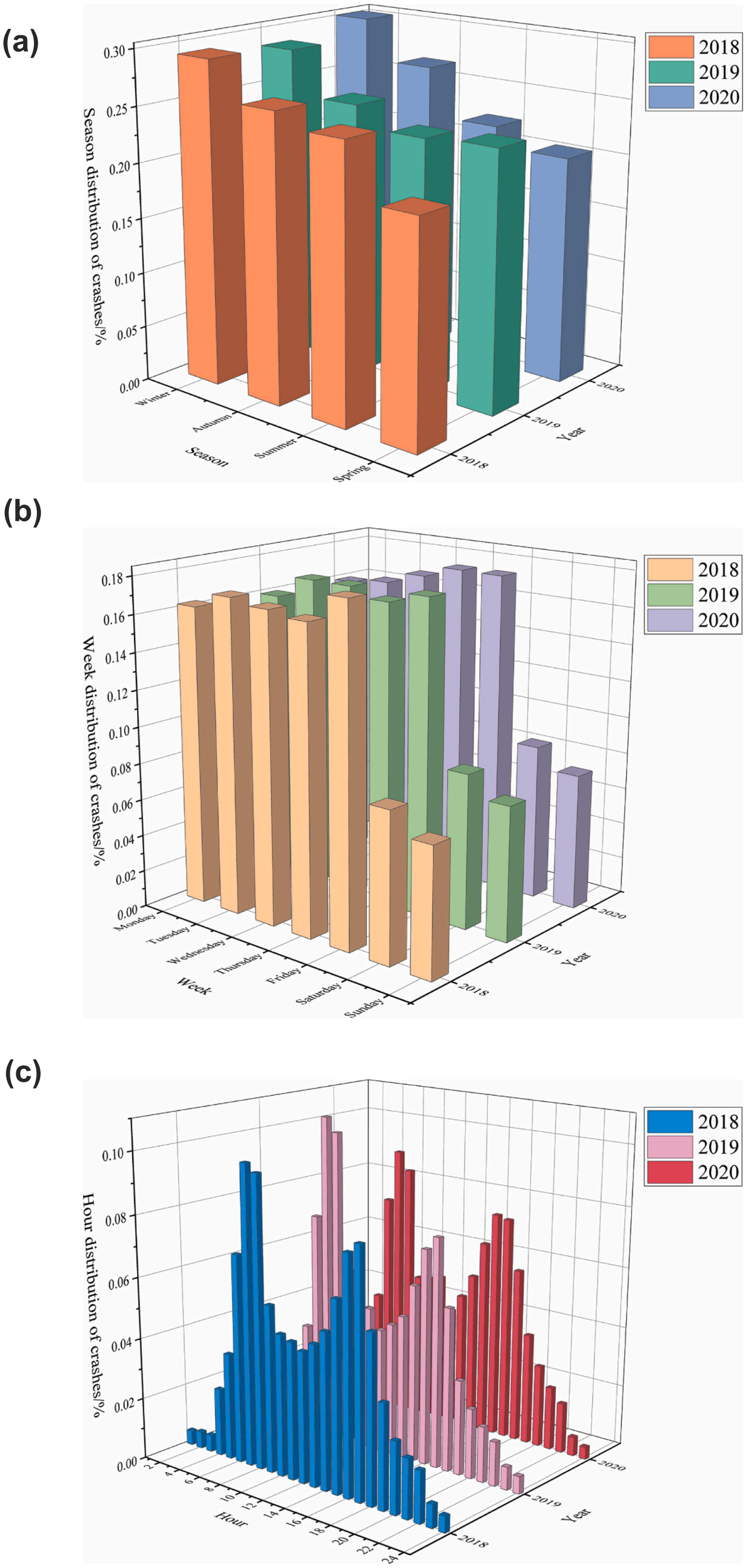


Fig. 1. Time distribution characteristics of motor vehicle traffic crashes.

continuous variables such as temperature, humidity, and population density, comprising a total of 19 attributes of information. Table 1 provides detailed information about the 19 attributes.

2.3. Data consolidation and processing

Regional population density and the level of GDP development may also be important factors affecting the severity of crashes, but few forecasts for motor vehicle traffic crash severity will consider these indicators. For this purpose, we gathered the population and GDP data for each state in the continental United States for the year 2020 and integrated it with the temporal information, location details, environmental conditions, and road conditions of the collision points using a GIS platform.

The severity of crashes in a dataset is categorized into four classes based on the degree of traffic delay: minor impact (short delay as a result of the accident), moderate impact, impactful, and major impact (long delay) [39,41,42]. The four severity classes exhibited a normal distribution, and there was a significant imbalance in the classification of crash severity data. To address this issue, the data can be balanced using oversampling or undersampling techniques. Undersampling aims to equalize the distribution of injury outcomes by randomly discarding samples in large categories, while oversampling increases the samples in small categories. Many researchers have found that oversampling methods, especially the Synthetic Minority Oversampling Technique (SMOTE), can significantly improve the accuracy of crash severity classification [29,43]. To enhance the model prediction accuracy and facilitate the visualization of the prediction results, the dataset undergoes a SMOTE to balance the data.

2.4. Spatial-temporal characterization of data structures

2.4.1. The temporal distribution of motor vehicle traffic crash severity

Motor vehicle traffic crashes are influenced by a variety of factors, such as road conditions, weather changes, and time of day. These factors have a direct impact on the speed of the vehicle and the traffic conditions, thereby affecting the likelihood of motor vehicle traffic crashes. For this reason, we analyzed the temporal characteristics of motor vehicle traffic crash severity in terms of seasons, days of the week, and times of the day. Fig. 1 illustrates the characteristics of the temporal distribution of motor vehicle traffic crash severity. The characteristics of the distribution from 2018 to 2020 show that motor vehicle traffic crashes are more likely to occur in winter, on weekdays, and during peak commuting hours. The lower temperatures in winter and the high levels of ice and snow on the roads result in reduced coefficients of friction, significantly increasing the likelihood of crashes. The rise in the number of vehicles on the road during weekdays results in increased traffic congestion, particularly during peak commuting hours when motor vehicle traffic crashes are more likely to happen. Moreover, the number of motor vehicle traffic crashes caused by driver fatigue starts to increase in the afternoon, and the severity of the crashes increases accordingly. Compared to weekdays, the incidence and motor vehicle traffic crash severity is lower on weekends because the number of vehicles on the road is relatively low. Motor vehicle traffic crashes are small

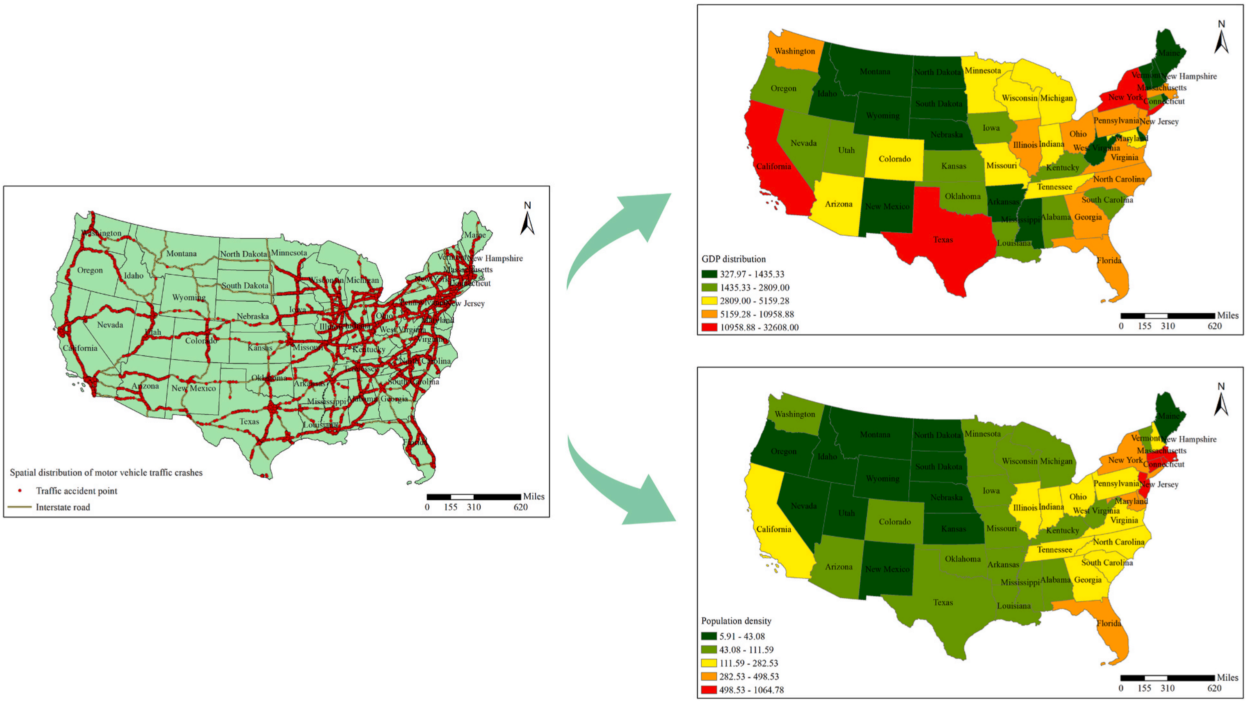


Fig. 2. Spatial distribution of motor vehicle traffic crashes compared with population density and GDP.

probability events that often suffer from zero-inflation issues, so data imbalance needs to be considered when developing predictive models. Although the number of interstate highway vehicle traffic crashes varies between different years, their temporal distribution characteristics remain largely consistent. The indicators in this study, including weather conditions, road factors, and temporal elements, are minimally affected by the Coronavirus disease 2019 (COVID-19), and societal factors such as population density are unlikely to undergo significant changes in the short term. Based on data availability and completeness, this study selects the 2020 dataset of interstate highway crashes in the contiguous United States as its research focus.

2.4.2. The spatial distribution of the motor vehicle traffic crash severity

Spatial heterogeneity and spatial correlation are important characteristics of the distribution of motor vehicle traffic crash severity. A comparison of the spatial distribution of motor vehicle traffic crashes with population density and GDP is shown in Fig. 2. The eastern, Great Lakes, and western coasts of the continental U.S. have a higher frequency of vehicle traffic crashes, while the central region has a lower frequency of crashes. There is a clear correlation between the occurrence of motor vehicle traffic crashes and the level of population density and GDP in the area.

The spatial-temporal statistical analysis of vehicle traffic crashes can provide a brief understanding of the spatial-temporal distribution characteristics of crashes. However, there exists a certain non-linear relationship between crashes influencing factors and the severity of vehicle traffic crashes. Therefore, further exploration is needed to uncover the intrinsic connection between the influencing factors and the severity of vehicle traffic crashes.

3. Methodology

The Random Forest algorithm is generally applicable to the prediction of motor vehicle traffic crash severity in terms of its nonlinear data processing capability, low noise dependence, and high accuracy [16,44,45] This section presents the implementation of the study and provides a summary of metaheuristic algorithms. Detailed explanations of the methods are found in some previous studies [46–54]. In addition to this, we introduce methods used to address imbalances in data structures and methods for evaluating model accuracy.

3.1. Research framework

17 impact indicators were selected as input values for predicting crash severity through a literature review combined with actual data. The data preprocessing was completed through data cleaning, merging, balancing, and normalization. To enhance the prediction accuracy of the model, the hyperparameters of the Random Forest were optimized using nine metaheuristic algorithms. Metaheuristic algorithms include Crested Porcupine Optimizer (CPO) [46], Horned Lizard Optimization Algorithm (HLOA) [47], Hippopotamus Optimization Algorithm (HO) [48], PID-based search algorithm (PID) [49], Triangulation Topology Aggregation Optimizer (TTAO) [50], Newton-Raphson-based optimizer (NBRO) [51], Football team training algorithm (FTTA) [52], Sparrow Search Algorithm (SSA) [53], and Dung Beetle Optimizer (DBO) [54]. Algorithms are compared with common classification prediction models such as Support Vector Machine (SVM) and BP Neural Network to ultimately determine the optimal prediction model. A GIS platform was utilized to

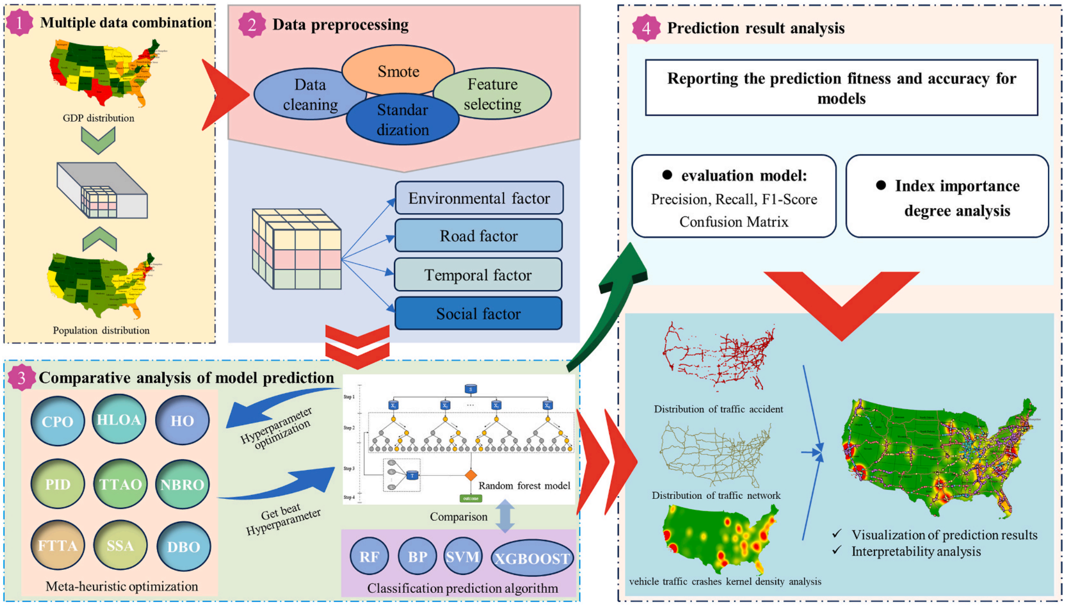

1 Multiple data combination
GDP distribution
Population distribution
2 Data preprocessing
Data cleaning
Smote
Standar dization
Feature selecting
Environmental factor
Road factor
Temporal factor
Social factor
3 Comparative analysis of model prediction
CPO
HLOA
HO
PID
TTAO
NBRO
FTTA
SSA
DBO
Meta-heuristic optimization
Hyperparameter optimization
Get beat Hyperparameter
Random forest model
Comparison
RF
BP
SVM
XGBOOST
Classification prediction algorithm
4 Prediction result analysis
Reporting the prediction fitness and accuracy for models
● evaluation model:
Precision, Recall, F1-Score
Confusion Matrix
● Index importance degree analysis
Distribution of traffic accident
Distribution of traffic network
vehicle traffic crashes kernel density analysis
✓ Visualization of prediction results
✓ Interpretability analysis

Fig. 3. The flowchart diagram of the current investigation.

visualize the prediction results, and crucial indicators were chosen for explanatory analysis based on their importance in the prediction. The research framework is shown in Fig. 3.

3.2. Random forest optimized by meta-heuristic algorithm

Random Forest [55] is a non-linear model proposed by Leo Breiman and Adele Cutler, adept at tackling regression and classification tasks (see Fig. 4). This method leverages bootstrap sampling to vary the training dataset, forming an ensemble of decision trees. In each tree's node training phase, features are randomly chosen without replacement from the entire set in a specified ratio. Throughout training, *k*-fold cross-validation is employed to mitigate overfitting.

In this study, we applied 4-fold cross-validation to the model for model training, where the filtered vehicle crash data were divided into four equal subsets. Among the four subsamples, three subsamples were used as the training data (75 %), and one subsample was used as the validation data (25 %). The *i-train* samples for RF are represented as $\mathrm{S} = [s_1, s_2, \ldots, s_i]$, and the *n-test* samples for RF are represented as $\mathrm{T} = [t_1, t_2, \ldots, t_n]$. A metaheuristic optimization algorithm is utilized to determine the number of decision trees, where each tree has a depth of d and employs M features at each node. The algorithm's flow can be summarized as follows.

(1) Training sets X are created by re-sampling the approach from the original data set i times. After a total of K rounds of extraction, K new sample sets are obtained: $(X_1, X_2, \ldots, X_K)$. The K sample sets that are independently sampled and have the same distribution will be used to generate K decision trees.
(2) Assuming there are M features in the feature space, during each round of generating the decision tree, m features (where $\mathrm{m} < \mathrm{M}$) are randomly selected from the feature space. Starting from the root node, a new feature set is formed from top to bottom, and a complete decision tree is generated. After K rounds of calculation, K decision trees are generated. K decision trees are combined to form a random forest. The fitness function for the meta-heuristic algorithm is generated during the random forest training process. Hyperparameter optimization for the random forest model is conducted using a meta-heuristic algorithm to determine the optimal hyperparameters.
(3) The samples $\mathrm{t_n}$ of the test set T are fed into a random forest containing optimal hyperparameters to enable decisions to be made on the test samples by each decision tree. Subsequently, the majority voting method is employed to vote on the decision results and determine the categorical prediction of motor vehicle traffic crash severity for $\mathrm{t_n}$.
(4) Step 3 is repeated n times until the classification prediction of the test sample T is completed.

3.3. Crested Porcupine Optimizer

Crown Porcupine Optimizer (CPO) [46] models various defensive behaviors observed in crown porcupines. These behaviors—sight, sound, scent, and physical attack—are ranked from least to most aggressive and are simulated across four distinct regions within CPO.

In the first region, CP employs the least aggressive defense strategy, being farther from the predator. Random values, generated

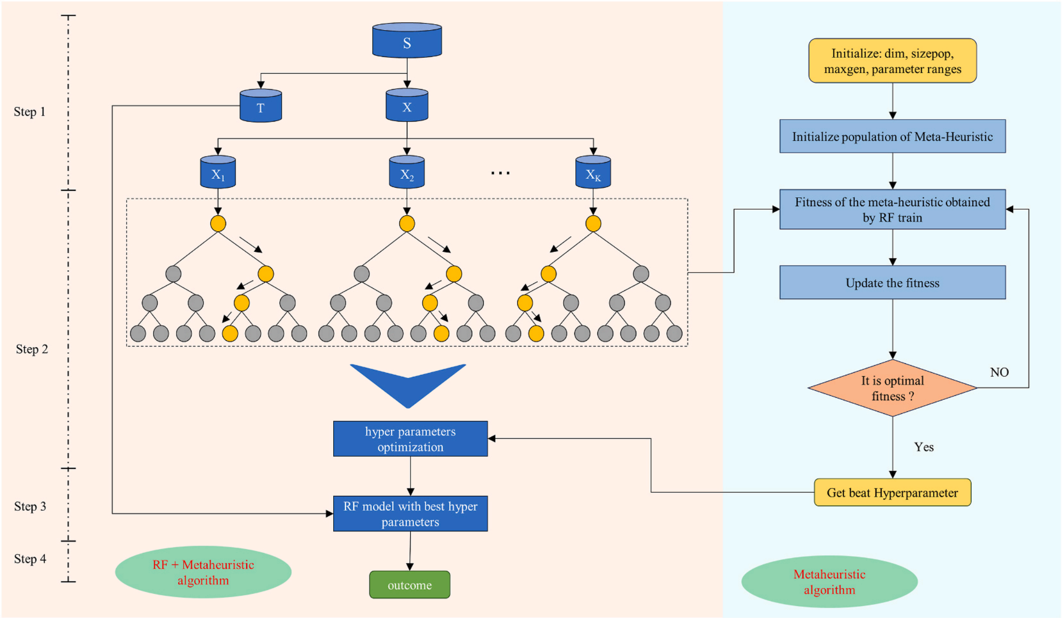

Step 1
Step 2
Step 3
Step 4
S
T
X
X_1
X_2
…
X_K
hyper parameters optimization
RF model with best hyper parameters
outcome
RF + Metaheuristic algorithm
Initialize: dim, sizepop, maxgen, parameter ranges
Initialize population of Meta-Heuristic
Fitness of the meta-heuristic obtained by RF train
Update the fitness
It is optimal fitness ?
NO
Yes
Get beat Hyperparameter
Metaheuristic algorithm

Fig. 4. Optimization of the random forest by meta-heuristic algorithm.

using a normal distribution, simulate these behaviors mathematically. Movement towards CPs is encouraged if these values fall outside the range of −1 to 1; otherwise, predators move away.

$$\overrightarrow{x_i^{t+1}} = \overrightarrow{x_i^t} + \tau_1 \times \left| 2 \times \tau_2 \times \overrightarrow{x_{CP}^t} - \overrightarrow{y_i^t} \right| \tag{1}$$

Where $\overrightarrow{x_i^t}$ is the position of the i-th individual at iteration t of the predator, $\overrightarrow{x_{CP}^t}$ is the best-obtained solution, $\overrightarrow{y_i^t}$ is the position of a predator at iteration t, τ_1 and τ_2 is a random value in the interval of [0, 1]. $\overrightarrow{y_i^t}$ is as follows:

$$\overrightarrow{y_i^t} = \frac{\overrightarrow{x_i^t} + \overrightarrow{x_r^t}}{2} \tag{2}$$

Where r is a random number between [1, N].

If the predator persists despite the initial defense strategy, a secondary defense mechanism is enacted. In the second strategy, the CP utilizes sound to intimidate the predator, increasing the intensity of its vocalizations as the predator draws nearer.

$$\overrightarrow{x_i^{t+1}} = \left(1 - \overrightarrow{U_1}\right) \times \overrightarrow{x_i^t} + \overrightarrow{U_1} \times \left(\overrightarrow{y} + \tau_3 \times \left(\overrightarrow{x_{r_1}^t} - \overrightarrow{x_{r_2}^t}\right)\right) \tag{3}$$

Where r_1 and r_2 are two random integers between [1, N], and τ_3 is a random value generated between 0 and 1. $\overrightarrow{U_1}$ is used to determine whether predators will continue to approach CP. $\overrightarrow{y}$ indicates the location of the predator.

If the predator continues to advance even after the second and third defense strategies are employed, the CP initiates a third defense mechanism.

$$\overrightarrow{x_i^{t+1}} = \left(1 - \overrightarrow{U_1}\right) \times \overrightarrow{x_i^t} + \overrightarrow{U_1} \times \left(\overrightarrow{x_{r_1}^t} + S_i^t \times \left(\overrightarrow{x_{r_2}^t} - \overrightarrow{x_{r_3}^t}\right) - \tau_3 \times \overrightarrow{\delta} \times \gamma_t \times S_i^t\right) \tag{4}$$

where r_3 is a random number between [1, N], $\overrightarrow{\delta}$ is a parameter used to control the search direction and is defined using Eq. (5), γ_t is the defense factor defined using Eq. (6), and S_i^t is the odor diffusion factor and is defined using Eq. (7).

$$\overrightarrow{\delta} = \begin{cases} +1\,, \text{ if } & \overrightarrow{rand} \leq 0.5 \\ -1\,, & Else \end{cases} \tag{5}$$

$$\gamma_t = 2 \times rand \times \left(1 - \frac{t}{t_{\max}}\right)^{\frac{t}{t_{\max}}} \tag{6}$$

$$S_i^t = \exp\left(\frac{f(x_i^t)}{\sum_{k=1}^{N} f(x_t^k) + \varepsilon}\right) \tag{7}$$

where $f(x_i^t)$ represents the objective function value of the i-th individual at iteration t, ε is a small value to avoid division by zero, $\overrightarrow{rand}$ is a vector including numerical values generated randomly between 0 and 1, *rand* is a variable including a number generated randomly between 0 and 1, N is the population size, t is the number of the current iteration, and t_{max} is the maximum number of iterations.

When all defense mechanisms fail, the prey attacks the predators, incapacitating or even killing them to protect themselves.

$$\overrightarrow{x_i^{t+1}} = \overrightarrow{x_{CP}^t} + (\alpha(1 - \tau_4) + \tau_4) \times \left(\delta \times \overrightarrow{x_{CP}^t} - \overrightarrow{x_i^t}\right) - \tau_5 \times \delta \times \gamma_t \times \overrightarrow{F_i^t} \tag{8}$$

where α is a convergence speed, τ_4, and τ_5 is a random value within the interval [0,1], and $\overrightarrow{F_i^t}$ is the average force of the CP that affected the i-th predator.

3.4. Evaluation metrics

Accuracy, precision, recall, F1-score, and confusion matrix are commonly used metrics to assess the performance of classification prediction models [56,57,].

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

Where TP (true positive) indicates proportion of actual positive cases that were correctly identified as positive by the model. FP (false positive) denotes samples that are actually in the negative category are incorrectly classified in the positive category. TN (true negative) signifies situations where the model correctly predicted negative outcomes. FN (false negative) refers Proportion of actual positive cases incorrectly identified as negative by the model.

Confusion matrices can be used to represent predictions using two-dimensional data. In this type of matrix, the x-axis displays the model's true labels, and the y-axis shows the predicted class labels. The confusion matrix allows us to visualize how the model classifies each class and observes incorrectly assigned to other classes.

3.5. Data balancing

Unbalanced data poses a significant challenge to prediction model accuracy, whereas crash severity classification typically adheres to a normal distribution. To tackle the imbalance in our dataset, both under-sampling and over-sampling techniques are commonly employed. In this study, we employ the Synthetic Minority Oversampling Technique (SMOTE) algorithm to address the issue. SMOTE oversamples minority class samples, ensuring an equal representation across all severity categories and enhancing the model's discriminative capability for each category in the imbalanced dataset. This approach helps to prevent redundancy in the sampled data, reduces overfitting risks, and improves the model's ability to generalize [18,58,59].

4. Results and discussions

In this section of the article, the prediction of motor vehicle traffic crash severity on interstate highways in the contiguous United States will be conducted using nine meta-heuristic algorithms (CPO, HLOA, HO, PID, TTAO, NBRO, FTTA, SSA, and DBO) combined with an RF model and a single model (RF, BP, SVM, and XGBoost). The predictive performance will be compared to determine the best classification prediction model. Finally, a GIS platform is used to conduct interpretive analysis of the prediction results and visualize them.

4.1. Model accuracy analysis

In this study, the dataset was divided into two parts: a training set consisting of 75 % and a test set containing 25 %. This division was conducted using stratified sampling to ensure equal representation of various graded crashes. Comparison criteria involved the precision, recall, F1-Score, overall accuracy, and confusion matrix of each model. And analyze the predictive accuracy of different gradations of motor vehicle traffic crash severity. The prediction results of the models regarding the accurate categorization of crash severity are shown in Table 2. Due to the large volume of data, random forest models optimized using nine meta-heuristic algorithms all achieve comparable prediction accuracies. The differences in accuracy among training sets are minimal, with an accuracy of 1.000 for each of the four levels of the training set when rounded to 0.001. However, the training set accuracies show varying differences. The three models selected as the most effective by comparing the precision of the models to show their test set confusion matrices are illustrated in Fig. 5(a–c), which compares the predicted results with the actual results.

Table 2 shows the effectiveness of the random forest optimized by meta-heuristic algorithm and the single classification prediction algorithm to predict the motor vehicle traffic crash severity. Among the single classification prediction models, RF has a good prediction effect, XGBoost has an average effect, while SVM and BP have poor prediction effects. The random forest optimized by meta-heuristic algorithm can predict motor vehicle traffic crash severity more effectively compared to a single classification prediction model. Among them, the CPO-RF model shows the best performance, with precision, recall, and F1 score exceeding 90 %. Meanwhile, the CPO-RF model has the best overall accuracy of 95.2 %. A 0.1 % variance in prediction accuracy can lead to the correct prediction of dozens or even hundreds of motor vehicle traffic crash samples due to the extensive dataset available. So, an increase in accuracy also means that more crashes can be predicted accurately.

The confusion matrix allows us to visualize how the model classifies each class and observes incorrectly assigned to other classes. The first three models with the highest accuracy were selected for analysis. In Fig. 5, it can be seen that class 1 and class 4 exhibit higher prediction accuracy, while class 2 and class 3 are more prone to incorrect predictions. The models can more accurately predict crashes of minor and significant impact, making it easier to anticipate the occurrence of more severe crashes.

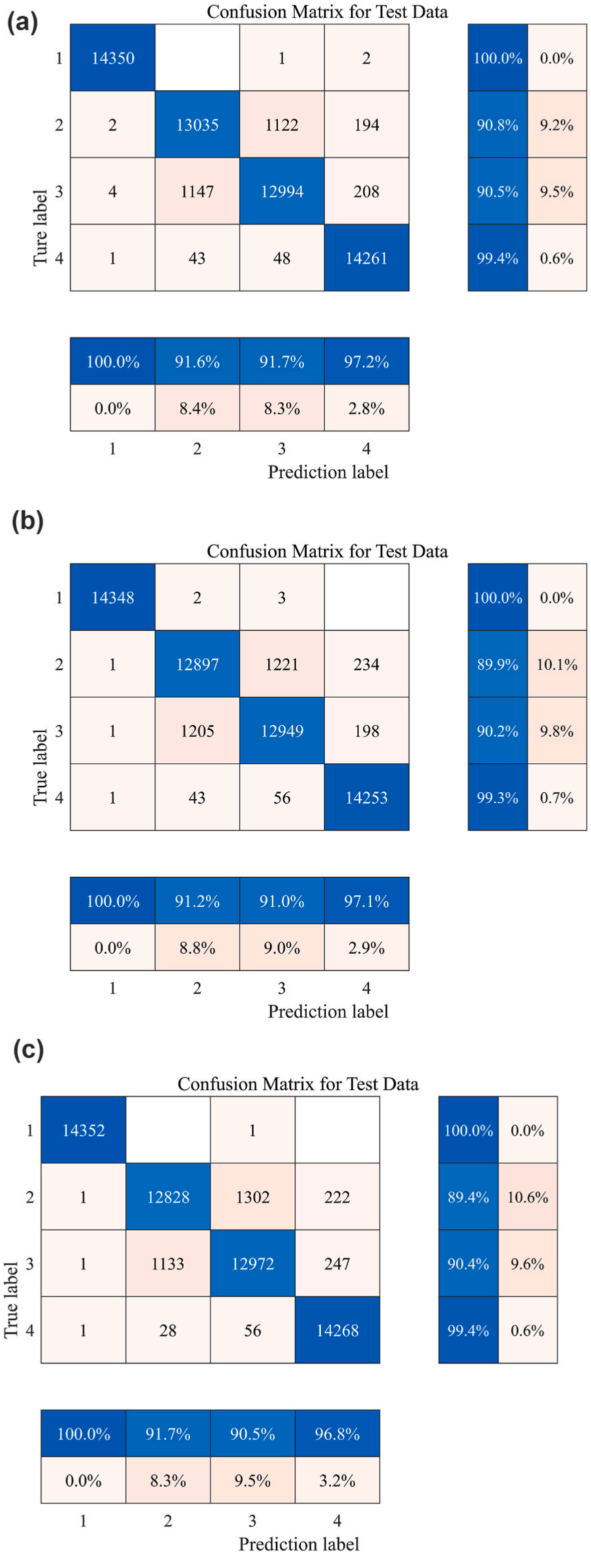
4.2. Importance of indicators

The three models with the best prediction results were selected to analyze the degree of importance of the indicators, and Fig. 6 shows the importance of the indicators in the prediction results of the three models CPO-RF, HO-RF, and DBO-RF, where 1–17 are the corresponding numbers of the indicators in Table 1.

Fig. 6 shows that the three model indicators with the best prediction effect have the same top six importance rankings. The six

Table 2
Model accuracy analysis.

Optimizer		Precision				Recall				F1 Score				Accuracy
		Class1	Class2	Class3	Class4	Class1	Class2	Class3	Class4	Class1	Class2	Class3	Class4	
CPO-RF	train	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	test	**1.000**	**0.916**	**0.917**	**0.972**	**1.000**	**0.908**	**0.905**	**0.994**	**1.000**	**0.908**	**0.907**	**0.983**	**0.952**
HLOA-RF	train	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	test	1.000	0.913	0.904	0.968	1.000	0.893	0.900	0.993	1.000	0.903	0.902	0.980	**0.946**
HO-RF	train	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	test	1.000	0.917	0.909	0.969	1.000	0.898	0.904	0.994	1.000	0.907	0.906	0.981	**0.949**
PID-RF	train	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	test	1.000	0.917	0.905	0.968	1.000	0.894	0.904	0.994	1.000	0.905	0.904	0.981	**0.948**
TTAO-RF	train	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	test	0.999	0.907	0.911	0.965	1.000	0.901	0.894	0.991	0.999	0.904	0.902	0.978	**0.946**
NBRO-RF	train	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	test	0.999	0.908	0.910	0.966	1.000	0.900	0.894	0.991	0.999	0.904	0.902	0.978	**0.946**
FTTA-RF	train	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	test	1.000	0.907	0.905	0.969	1.000	0.894	0.898	0.990	1.000	0.900	0.901	0.979	**0.945**
SSA-RF	train	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	test	1.000	0.909	0.907	0.965	1.000	0.897	0.894	0.991	1.000	0.903	0.900	0.978	**0.945**
DBO-RF	train	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	test	1.000	0.912	0.910	0.971	1.000	0.899	0.902	0.993	1.000	0.905	0.904	0.981	**0.948**
RF	train	1.000	0.952	0.938	0.968	1.000	0.925	0.940	0.994	1.000	0.938	0.939	0.981	0.965
	test	1.000	0.908	0.893	0.949	1.000	0.871	0.890	0.990	1.000	0.889	0.891	0.969	**0.937**
XGBoost	train	0.997	0.886	0.684	0.953	1.000	0.733	0.879	0.844	1.000	0.938	0.939	0.981	0.864
	test	0.995	0.880	0.671	0.947	1.000	0.724	0.869	0.829	1.000	0.889	0.891	0.969	**0.856**
SVM	train	0.781	0.669	0.655	0.672	0.993	0.552	0.520	0.745	0.874	0.605	0.580	0.707	0.703
	test	0.782	0.671	0.656	0.667	0.992	0.552	0.522	0.745	0.875	0.606	0.581	0.704	**0.702**
BP	train	0.980	0.744	0.667	0.807	1.000	0.654	0.762	0.781	0.990	0.696	0.711	0.794	0.800
	test	0.782	0.671	0.656	0.667	0.992	0.552	0.522	0.745	0.875	0.606	0.581	0.704	**0.798**



(a)

	1	2	3	4		
1	14350		1	2	100.0%	0.0%
2	2	13035	1122	194	90.8%	9.2%
3	4	1147	12994	208	90.5%	9.5%
4	1	43	48	14261	99.4%	0.6%
	100.0%	91.6%	91.7%	97.2%		
	0.0%	8.4%	8.3%	2.8%		

(b)

	1	2	3	4		
1	14348	2	3		100.0%	0.0%
2	1	12897	1221	234	89.9%	10.1%
3	1	1205	12949	198	90.2%	9.8%
4	1	43	56	14253	99.3%	0.7%
	100.0%	91.2%	91.0%	97.1%		
	0.0%	8.8%	9.0%	2.9%		

(c)

	1	2	3	4		
1	14352		1		100.0%	0.0%
2	1	12828	1302	222	89.4%	10.6%
3	1	1133	12972	247	90.4%	9.6%
4	1	28	56	14268	99.4%	0.6%
	100.0%	91.7%	90.5%	96.8%		
	0.0%	8.3%	9.5%	3.2%		

Fig. 5. The confusion matrix generated from the prediction of the test data.

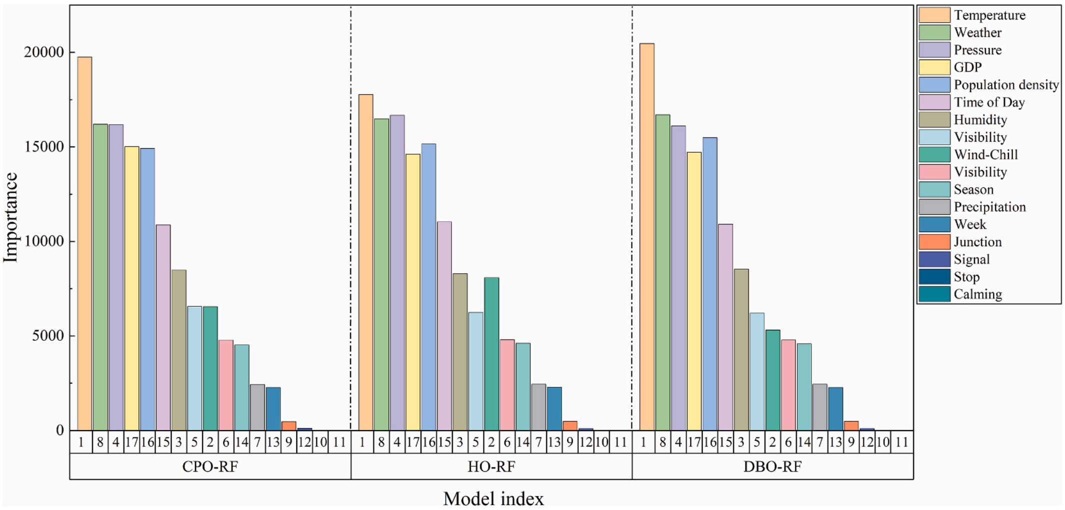


Fig. 6. Ranking of importance degree of indicators in CPO-RF, HO-RF, and DBO-RF model.

indicators that have the greatest impact on motor vehicle traffic crashes are 1 (Temperature), 8 (Weather), 4 (Pressure), 17 (GDP), 16 (Population density), and 15 (Time of Day).

4.3. Results interpretive

The CPO-RF model with the best performance is chosen for visualizing and interpreting the prediction results. The visualization part involves spatially representing predicted crash points using a GIS platform. In this part of the study, the test set and prediction results of the model are first processed with Inverse SMOTE to obtain the actual crash points. We found that the prediction accuracy of the data after the Inverse SMOTE processing reached 99.6 %, and the confusion matrix after the Inverse SMOTE processing is shown in Fig. 7.

4.3.1. Analysis of results

The location information of the data is utilized to visualize the prediction results on the GIS platform. Since there is only one sample in class 1 and it is predicted correctly, spatial visualization analysis is conducted only for class 2, class 3, and class 4, as shown in Fig. 8. The incorrectly predicted points are mainly located at two major transportation intersections in the California region.

4.3.2. Interpretive analysis of results

Environmental indicator 1 (Temperature), Time indicator 5 (Time of day), and Social indicator 17 (GDP) were selected for the

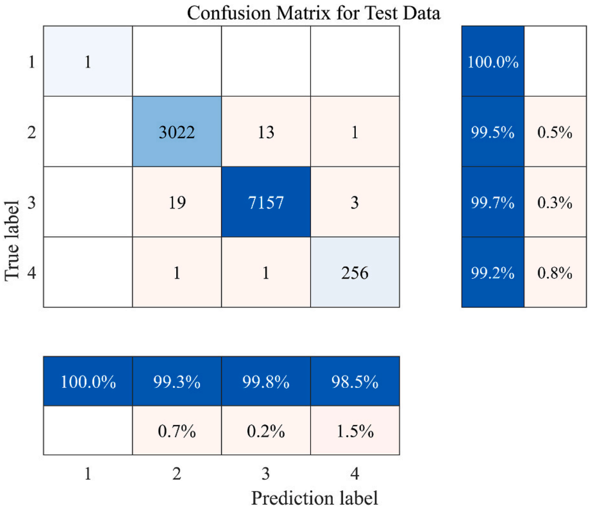


Fig. 7. Confusion matrix of predicted results after inverse Smote processing.

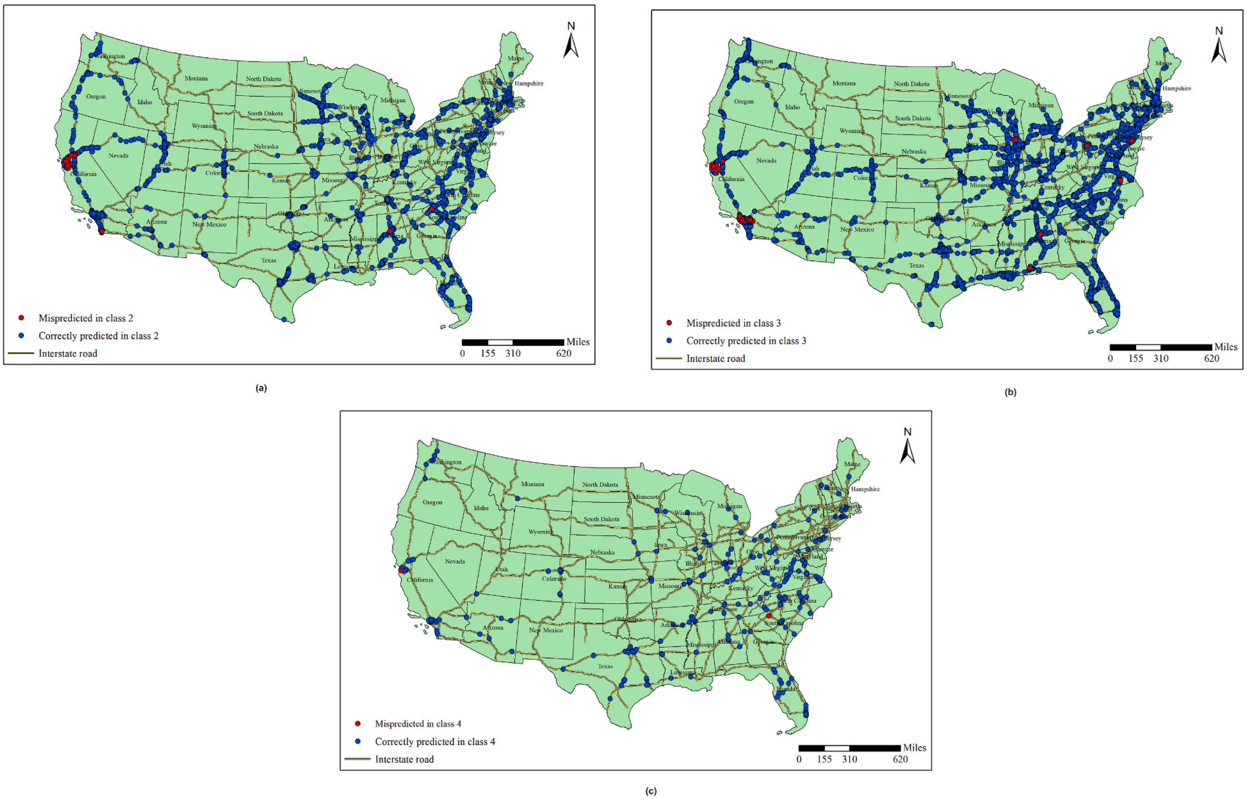


Fig. 8. Visualization of CPO-RF model prediction results.

explanatory analysis of the predictor indicators based on the top six rankings of the indicators' importance in Section 4.2. For comparison, the continuous variables 1 (Temperature) and 17 (GDP) were categorized into four classes using the quantile classification method in GIS, while the categorical variable 15 (Time of day) is presented in Table 1. Kernel density analysis of motor vehicle traffic crash points using the spatial density analysis function in GIS, as shown in Fig. 9. The areas with frequent motor vehicle traffic crashes are primarily concentrated in densely populated areas, intersections, and other locations that may lead to complex road conditions. The six indicators are combined with kernel density analysis for interpretive analysis, as shown in Fig. 10.

Fig. 10(a) illustrates that the crash points are mainly concentrated in the high-risk area of motor vehicle traffic crashes, with most of them occurring in high-temperature locations. This indicates a clear positive correlation between motor vehicle traffic crashes and temperature, although it is not a straightforward linear relationship. The rise in temperature causes the ground temperature to increase, resulting in a decrease in asphalt pavement friction, which can easily lead to motor vehicle traffic crashes [60]. Furthermore, high temperatures also have a significant impact on human emotions [61].

GDP is an important factor influencing the number of national private vehicles and traffic conditions. Relevant studies have found that in regions with higher levels of socio-economic development people's willingness to travel increases and commercial vehicle movements will also increase, which will increase the risk of traffic vehicle traffic crashes. In regions with lower levels of economic development, factors such as poorer vehicle levels, more old vehicles on the roads, and poorer road infrastructure will also affect vehicle traffic safety. Mobility may be the most critical determinant of vehicle traffic safety [62]. Fig. 10(b) illustrates that California and Texas have higher levels of GDP development, and simultaneously, they are also high-incidence areas for vehicle traffic crashes. The central region of the U.S. mainland has a lower level of GDP development, and simultaneously, there are fewer vehicle traffic crashes. It is worth noting that vehicle traffic crashes have become increasingly common in some states with lower levels of GDP development, such as Minnesota, Tennessee, and Alabama. This trend reflects the quality of traffic management in these regions.

Time of Day is also an important factor that affects vehicle traffic crashes. Fig. 10(c) shows that the incidence of vehicle traffic crashes is greatly affected by the commuting peak in the states of Utah, Texas, and New Hampshire. It is noteworthy that motor vehicle traffic crashes in California and Washington are largely unaffected by the commuting peak, which may be related to the level of public transportation development in the cities [63]. In addition, most vehicle traffic crashes in Illinois occur at night, highlighting the need for additional traffic facilities to ensure safe driving, such as increased lighting, speed limit signs, and other measures.

5. Conclusion

Motor vehicle traffic crash severity prediction is crucial for preventing future crashes and mitigating different types of motor

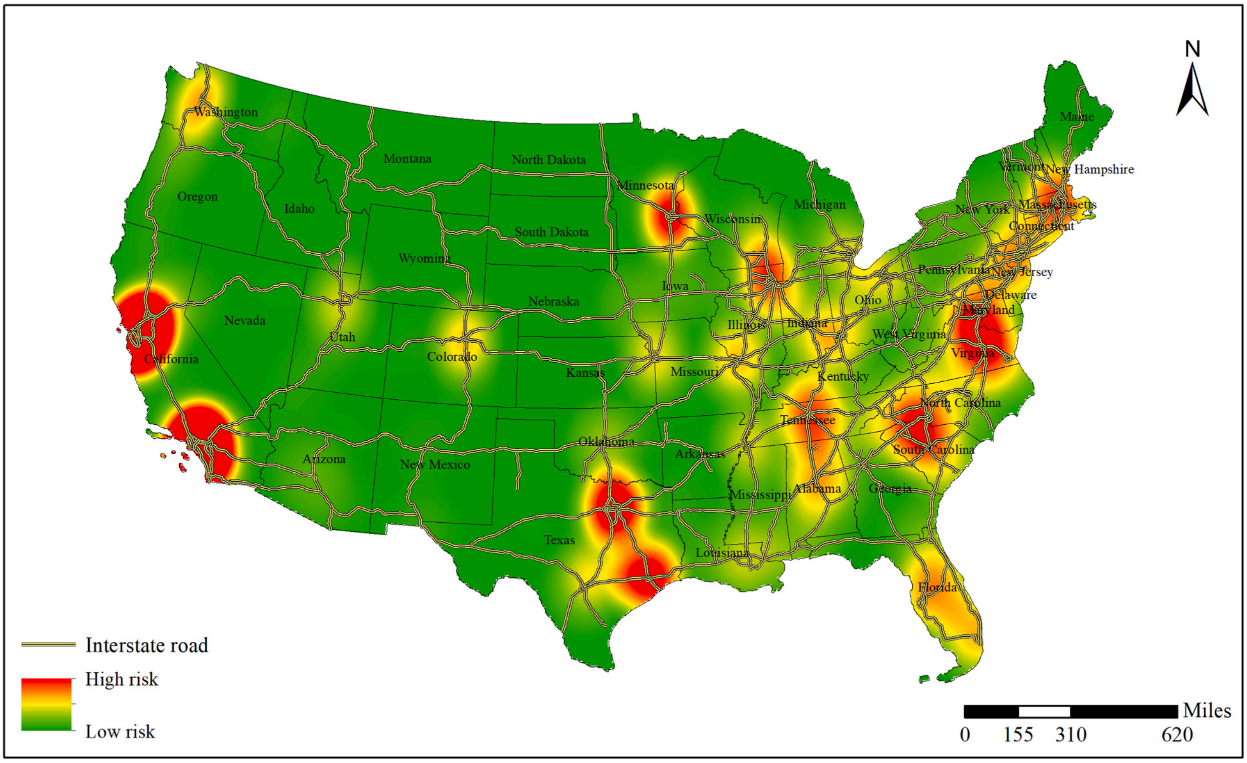


Fig. 9. Kernel density analysis of motor vehicle traffic crashes locations.

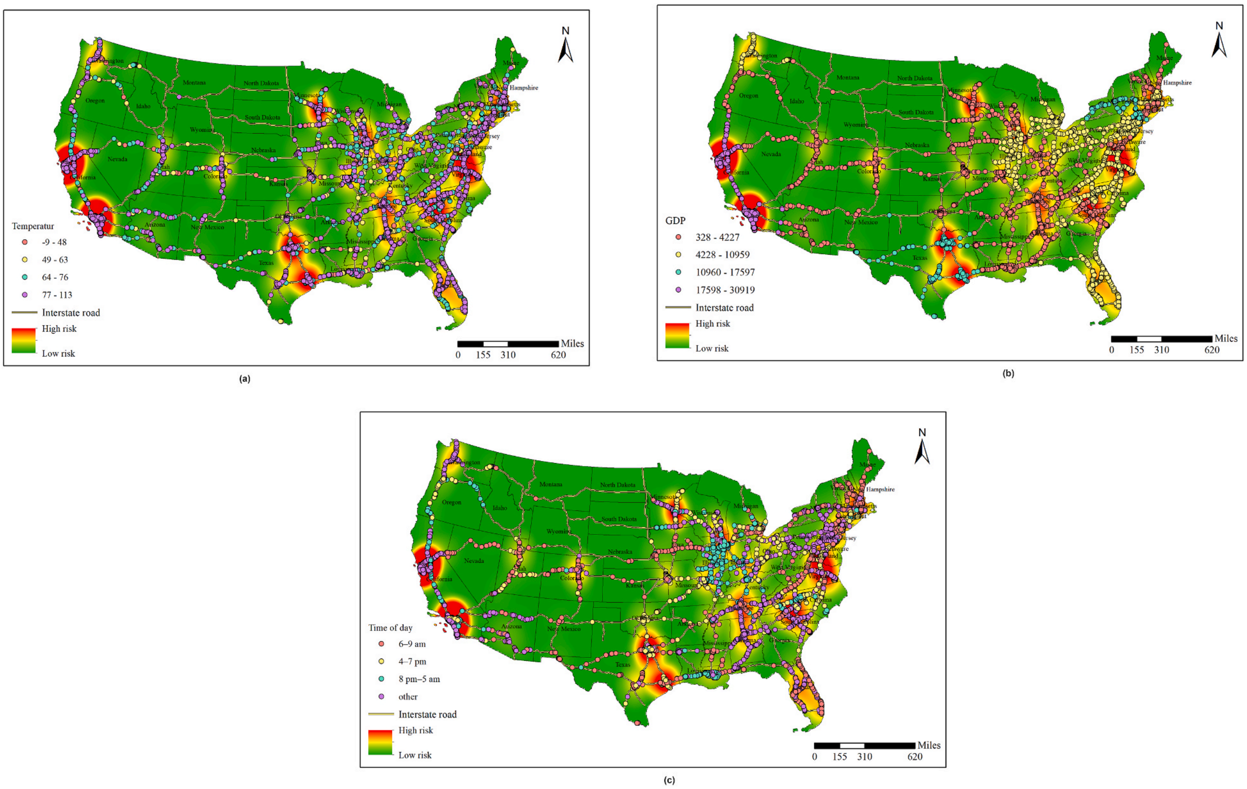


Fig. 10. Important Impact Indicator Analysis of Motor vehicle traffic crashes.

vehicle traffic crash severity. In this study, we design an analytical framework for predicting motor vehicle traffic crashes using multiple meta-inspired optimization algorithms combined with random forests. Utilizing the framework, predict the severity of motor vehicle traffic crashes based on the 2020 motor vehicle traffic crash statistics in the continental United States. This framework mainly includes the comparative analysis of combined optimization algorithms and single algorithms, as well as the information fusion and visualization on a GIS platform.

The CPO-RF model produces the most accurate predictions compared to a single model and random forest optimized by other meta-heuristic algorithms. It can achieve precision, recall, and F1 scores above 90 % for various severity classes, with an overall accuracy of 95.2 %. The CPO-RF model demonstrates better overall performance compared to other combination models and single classification prediction models. After applying the inverse SMOTE treatment to the prediction results of the CPO-RF model, we found that the prediction accuracy increased to 99.6 %. The three model indicators with the best prediction effect have the same top 6 importance rankings. The six indicators that have the greatest impact on motor vehicle traffic crashes are 1 (Temperature), 8 (Weather), 4 (Pressure), 17 (GDP), 16 (Population Density), and 15 (Time of Day).

GIS was used to visualize the prediction results, pinpointing the specific locations of the inaccurately predicted crash points. Through interpretive analysis of key indicators of motor vehicle traffic crash severity, we can offer practical suggestions to urban planners for various regions. Studies show that there is a positive correlation between motor vehicle traffic crashes and temperature and GDP. It is important to consider the influence of temperature and GDP on motor vehicle traffic crashes when implementing traffic management strategies. Some regions, like Minnesota and Tennessee, need to enhance their traffic management. For example, speed limits could be imposed and additional road traffic monitoring devices installed to ensure safe driving. Furthermore, enhancing traffic morality and safety consciousness among residents in the area is crucial. When necessary, strict legal regulations on driving behavior could be implemented to bolster traffic legal awareness. Regarding periods within a day, states such as Utah and Texas exhibit a higher incidence of traffic crashes influenced by peak commuting hours. However, California and Washington benefit from well-developed public transportation systems, resulting in a vehicle traffic crash rate that is largely unaffected by commuting peaks. Illinois should consider implementing traffic facilities such as nighttime lighting installations and fatigue driving warning systems to ensure the safety of nighttime vehicle operations. City managers can use our model and its interpretability to predict vehicle traffic crashes more accurately. They can also identify factors that contribute to motor vehicle traffic crash severity and choose effective measures to address issues in transportation and urban development.

However, due to the complexity of motor vehicle traffic crashes, our study has some limitations. For example, the original data collection measures are limited, and indicators such as drivers and road alignment design are not included. Because factor selection is a systematic process, despite some experiences being reported in previous references, further study is necessary to select a more reasonable set of factors.

Data availability statement

The data used in this study is available at reasonable request from the corresponding author.

CRediT authorship contribution statement

Xing Wang: Writing – original draft, Methodology, Conceptualization. **Yikun Su:** Project administration, Funding acquisition. **Zhizhe Zheng:** Data curation. **Liang Xu:** Software, Resources.

Declaration of competing interest

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

Acknowledgment

The authors greatly appreciate the financial support of the Department of Transportation of Guangxi Province (Project No. 2019-ZD5-028).

References

[1] National Center for Statistics and Analysis, Overview of the 2020 Crash Investigation Sampling System (Traffic Safety Facts Research Note. Report No. DOT HS 813 255), National Highway Traffic Safety Administration, 2022, February.

[2] M. Burzynska, M. Pikala, Decreasing trends in road traffic mortality in Poland: a twenty-year analysis, Int. J. Environ. Res. Publ. Health 18 (2021) 10411, https://doi.org/10.3390/ijerph181910411.

[3] M.A. Aga, B.T. Woldeamanuel, M. Tadesse, Statistical modeling of numbers of human deaths per road traffic accident in the Oromia region, Ethiopia, PLoS One 16 (2021) e0251492, https://doi.org/10.1371/journal.pone.0251492.

[4] B. Xie, Z. An, Y. Zheng, Z. Li, Incorporating transportation safety into land use planning: pre-assessment of land use conversion effects on severe crashes in urban China, Appl. Geogr. 103 (2019) 1–11, https://doi.org/10.1016/j.apgeog.2018.12.003.

[5] T. Chen, N.N. Sze, S. Chen, S. Labi, Q. Zeng, Analysing the main and interaction effects of commercial vehicle mix and roadway attributes on crash rates using a Bayesian random-parameter Tobit model, Accid. Anal. Prev. 154 (2021) 106089, https://doi.org/10.1016/j.aap.2021.106089.

[6] National Center for Statistics and Analysis, Overview of Motor Vehicle Traffic Crashes in 2022 (Traffic Safety Facts Research Note. Report No. DOT HS 813 560), National Highway Traffic Safety Administration, 2024, April.

[7] Y.-R. Shiau, C.-H. Tsai, Y.-H. Hung, Y.-T. Kuo, The application of data mining technology to build a forecasting model for classification of road traffic accidents, Math. Probl Eng. 2015 (2015) 170635, https://doi.org/10.1155/2015/170635.

[8] S. Mokhtarimousavi, J.C. Anderson, A. Azizinamini, M. Hadi, Improved support vector machine models for work zone crash injury severity prediction and analysis, Transport. Res. Rec.: J. Transport. Res. Board 2673 (11) (2019) 680–692, https://doi.org/10.1177/0361198119845899.

[9] S. Mokhtarimousavi, J.C. Anderson, A. Azizinamini, M. Hadi, Factors affecting injury severity in vehicle-pedestrian crashes: a day-of-week analysis using random parameter ordered response models and Artificial Neural Networks, International Journal of Transportation Science and Technology 9 (2) (2020) 100–115, https://doi.org/10.1016/j.ijtst.2020.01.001.

[10] W. Zhang, Z. Zhou, L. Li, R. Huang, Identifying significant injury severity risk factors in traffic accidents based on the machine learning methods, in: CICTP 2019: Transportation in China - Connecting the World - Proceedings of the 19th COTA International Conference of Transportation Professionals, American Society of Civil Engineers (ASCE), 2019, pp. 3759–3770, https://doi.org/10.1061/9780784482292.326.

[11] Q. Zeng, W. Hao, J. Lee, F. Chen, Investigating the impacts of real-time weather conditions on freeway crash severity: a Bayesian spatial analysis, Int. J. Environ. Res. Publ. Health 17 (8) (2020) 2768, https://doi.org/10.3390/ijerph17082768.

[12] G. Pillajo-Quijia, B. Arenas-Ramírez, C. González-Fernández, F. Aparicio-Izquierdo, Influential factors on injury severity for drivers of light trucks and vans with machine learning methods, Sustainability 12 (4) (2020) 1324, https://doi.org/10.3390/su12041324.

[13] A.K. Panicker, G. Ramadurai, Injury severity prediction model for two-wheeler crashes at mid-block road sections, Int. J. Crashworthiness 27 (2) (2022) 328–336, https://doi.org/10.1080/13588265.2020.1806644.

[14] J. Lee, T. Yoon, S. Kwon, J. Lee, Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: seoul city study, Appl. Sci. 10 (1) (2020) 129, https://doi.org/10.3390/app10010129.

[15] N.S. Hadjidimitriou, M. Lippi, M. Dell'Amico, A. Skiera, Machine learning for severity classification of accidents involving powered two wheelers, IEEE Trans. Intell. Transport. Syst. 21 (10) (2019) 4308–4317, https://doi.org/10.1109/TITS.2019.2939624.

[16] I.C. Obasi, C. Benson, Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents, Heliyon 9 (8) (2023) e18812, https://doi.org/10.1016/j.heliyon.2023.e18812.

[17] X. Shen, S. Wei, Application of XGBoost for hazardous material road transport accident severity analysis, IEEE Access 8 (2020) 206806–206819, https://doi.org/10.1109/ACCESS.2020.3037922.

[18] X. Wen, Y. Xie, L. Jiang, Z. Pu, T. Ge, Applications of machine learning methods in traffic crash severity modelling: current status and future directions, Transport Rev. 41 (6) (2021) 855–879, https://doi.org/10.1080/01441647.2021.1954108.

[19] Z. Li, P. Liu, W. Wang, C. Xu, Using support vector machine models for crash injury severity analysis, Accid. Anal. Prev. 45 (2012) 478–486, https://doi.org/10.1016/j.aap.2011.08.016.

[20] J. Wang, Y. Kong, T. Fu, Expressway crash risk prediction using back propagation neural network: a brief investigation on safety resilience, Accid. Anal. Prev. 124 (2019) 180–192, https://doi.org/10.1016/j.aap.2019.01.007.

[21] C. Gutierrez-Osorio, C. Pedraza, Modern data sources and techniques for analysis and forecast of road accidents: a review, J. Traffic Transport. Eng. 7 (4) (2020) 432–446, https://doi.org/10.1016/j.jtte.2020.05.002.

[22] Y. Yang, K. Wang, Z. Yuan, D. Liu, Predicting freeway traffic crash severity using XGBoost-Bayesian network model with consideration of features interaction, J. Adv. Transport. 2022 (2022), https://doi.org/10.1155/2022/4257865.

[23] M. Yan, Y. Shen, Traffic accident severity prediction based on random forest, Sustainability 14 (3) (2022) 1729, https://doi.org/10.3390/su14031729.

[24] M.A. Rahim, H.M. Hassan, A deep learning based traffic crash severity prediction framework, Accid. Anal. Prev. 154 (2021) 106090, https://doi.org/10.1016/j.aap.2021.106090.

[25] Z. Yang, W. Zhang, J. Feng, Predicting multiple types of traffic accident severity with explanations: a multi-task deep learning framework, Saf. Sci. 146 (2022) 105522, https://doi.org/10.1016/j.ssci.2021.105522.

[26] A. Ziakopoulos, A. Kontaxi, G. Yannis, Analysis of mobile phone use engagement during naturalistic driving through explainable imbalanced machine learning, Accid. Anal. Prev. 181 (2023) 106936, https://doi.org/10.1016/j.aap.2022.106936.

[27] C. Ma, Y. Peng, L. Wu, X. Guo, X. Wang, X. Kong, Application of machine learning techniques to predict the occurrence of distraction-affected crashes with phone-use data, Transport. Res. Rec. 2676 (2) (2022) 692–705, https://doi.org/10.1177/03611981211045371.

[28] A.B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, A.K. Mohammadian, Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis, Accid. Anal. Prev. 136 (2020) 105405, https://doi.org/10.1016/j.aap.2019.105405.

[29] J. Ma, Y. Ding, J.C. Cheng, Y. Tan, V.J. Gan, J. Zhang, Analyzing the leading causes of traffic fatalities using XGBoost and grid-based analysis: a city management perspective, IEEE Access 7 (2019) 148059–148072, https://doi.org/10.1109/ACCESS.2019.2946401.

[30] M. Umer, S. Sadiq, A. Ishaq, S. Ullah, N. Saher, H.A. Madni, Comparison analysis of tree based and ensembled regression algorithms for traffic accident severity prediction, arXiv preprint arXiv:2010.14921 (2020), https://doi.org/10.48550/arXiv.2010.14921.

[31] S. Ramya, S.K. Reshma, V.D. Manogna, Y.S. Saroja, G.S. Gandhi, Accident severity prediction using data mining methods, International Journal of Scientific Research in Computer Science, Engineering and Information Technology 5 (2) (2019) 528–536, https://doi.org/10.32628/CSEIT195293.

[32] A. Ziakopoulos, G. Yannis, A review of spatial approaches in road safety, Accid. Anal. Prev. 135 (2020) 105323, https://doi.org/10.1016/j.aap.2019.105323.

[33] A. Loukaitou-Sideris, R. Liggett, H.G. Sung, Death on the crosswalk: a study of pedestrian-automobile collisions in Los Angeles, J. Plann. Educ. Res. 26 (3) (2007) 338–351, https://doi.org/10.1177/0739456x06297008.

[34] S. Mohammed, A.H. Alkhereibi, A. Abulibdeh, R.N. Jawarneh, P. Balakrishnan, GIS-based spatiotemporal analysis for road traffic crashes; in support of sustainable transportation Planning, Transp. Res. Interdiscip. Perspect. 20 (2023) 100836, https://doi.org/10.1016/j.trip.2023.100836.

[35] Y. Li, M. Abdel-Aty, J. Yuan, Z. Cheng, J. Lu, Analyzing traffic violation behavior at urban intersections: a spatio-temporal kernel density estimation approach using automated enforcement system data, Accid. Anal. Prev. 141 (2020) 105509, https://doi.org/10.1016/j.aap.2020.105509.

[36] N.H.M. Aghasi, Introducing GIS as legitimate instrument to deal with road accident data: a case study of Iran, Tehran, Spatial information research 25 (1) (2017) 151–159, https://doi.org/10.1007/s41324-017-0083-9.

[37] L. Hu, X. Wu, J. Huang, Y. Peng, W. Liu, Investigation of clusters and injuries in pedestrian crashes using GIS in Changsha, China, Saf. Sci. 127 (2020) 104710, https://doi.org/10.1016/j.ssci.2020.104710.

[38] F. Jiang, K.K.R. Yuen, E.W.M. Lee, Analysis of motorcycle accidents using association rule mining-based framework with parameter optimization and GIS technology, J. Saf. Res. 75 (2020) 292–309, https://doi.org/10.1016/j.jsr.2020.09.004.

[39] US Accidents, 2016 - 2023, https://doi.org/10.34740/kaggle/ds/199387.

[40] H. Jeong, Y. Jang, P.J. Bowman, N. Masoud, Classification of motor vehicle crash injury severity: a hybrid approach for imbalanced data, Accid. Anal. Prev. 120 (2018) 250–261, https://doi.org/10.1016/j.aap.2018.08.025.

[41] S. Moosavi, M.H. Samavatian, S. Parthasarathy, R. Ramnath, A countrywide traffic accident dataset (2019), https://doi.org/10.48550/arXiv.1906.05409 arXiv preprint arXiv:1906.05409.

[42] S. Moosavi, M.H. Samavatian, S. Parthasarathy, R. Teodorescu, R. Ramnath, Accident risk prediction based on heterogeneous sparse data: new dataset and insights, in: Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2019, November, pp. 33–42, https://doi.org/10.1145/3347146.3359078.

[43] R.O. Mujalli, L. Garach, G. López, T. Al-Rousan, Evaluation of injury severity for pedestrian–vehicle crashes in Jordan using extracted rules, J. Transport. Eng., Part A: Systems 145 (7) (2019) 04019028, https://doi.org/10.1061/JTEPBS.0000244.

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

[45] K. Santos, J.P. Dias, C. Amado, A literature review of machine learning algorithms for crash injury severity prediction, J. Saf. Res. 80 (2022) 254–269, https://doi.org/10.1016/j.jsr.2021.12.007.

[46] M. Abdel-Basset, R. Mohamed, M. Abouhawwash, Crested Porcupine Optimizer: a new nature-inspired metaheuristic, Knowl. Base Syst. 284 (2024) 111257, https://doi.org/10.1016/j.knosys.2023.111257.

[47] H. Peraza-Vázquez, A. Peña-Delgado, M. Merino-Treviño, A.B. Morales-Cepeda, N. Sinha, A novel metaheuristic inspired by horned lizard defense tactics, Artif. Intell. Rev. 57 (3) (2024) 59, https://doi.org/10.1007/s10462-023-10653-7.

[48] M.H. Amiri, N. Mehrabi Hashjin, M. Montazeri, S. Mirjalili, N. Khodadadi, Hippopotamus optimization algorithm: a novel nature-inspired optimization algorithm, Sci. Rep. 14 (1) (2024) 5032, https://doi.org/10.1038/s41598-024-54910-3.

[49] Y. Gao, PID-based search algorithm: a novel metaheuristic algorithm based on PID algorithm, Expert Syst. Appl. 232 (2023) 120886, https://doi.org/10.1016/j.eswa.2023.120886.

[50] S. Zhao, T. Zhang, L. Cai, R. Yang, Triangulation topology aggregation optimizer: a novel mathematics-based meta-heuristic algorithm for continuous optimization and engineering applications, Expert Syst. Appl. 238 (2024) 121744, https://doi.org/10.1016/j.eswa.2023.121744.

[51] R. Sowmya, M. Premkumar, P. Jangir, Newton-Raphson-based optimizer: a new population-based metaheuristic algorithm for continuous optimization problems, Eng. Appl. Artif. Intell. 128 (2024) 107532, https://doi.org/10.1016/j.engappai.2023.107532.

[52] Z. Tian, M. Gai, Football team training algorithm: a novel sport-inspired meta-heuristic optimization algorithm for global optimization, Expert Syst. Appl. 245 (2024) 123088, https://doi.org/10.1016/j.eswa.2023.123088.

[53] J. Xue, B. Shen, A novel swarm intelligence optimization approach: sparrow search algorithm, Systems science & control engineering 8 (1) (2020) 22–34, https://doi.org/10.1080/21642583.2019.1708830.

[54] J. Xue, B. Shen, Dung beetle optimizer: a new meta-heuristic algorithm for global optimization, J. Supercomput. 79 (7) (2023) 7305–7336, https://doi.org/10.1007/s11227-022-04959-6.

[55] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[56] L. Pérez-Sala, M. Curado, L. Tortosa, J.F. Vicent, Deep learning model of convolutional neural networks powered by a genetic algorithm for prevention of traffic accidents severity, Chaos, Solit. Fractals 169 (2023) 113245, https://doi.org/10.1016/j.chaos.2023.113245.

[57] L. Yu, B. Du, X. Hu, L. Sun, L. Han, W. Lv, Deep spatio-temporal graph convolutional network for traffic accident prediction, Neurocomputing 423 (2021) 135–147, https://doi.org/10.1016/j.neucom.2020.09.043.

[58] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357, https://doi.org/10.1613/jair.953.

[59] K. Li, H. Xu, X. Liu, Analysis and visualization of accidents severity based on LightGBM-TPE, Chaos, Solit. Fractals 157 (2022) 111987, https://doi.org/10.1016/j.chaos.2022.111987.

[60] X. Zhao, A. Shen, B. Ma, Temperature adaptability of asphalt pavement to high temperatures and significant temperature differences, Adv. Mater. Sci. Eng. 2018 (1) (2018) 9436321, https://doi.org/10.1155/2018/9436321.

[61] S. Xianglong, Z. Hu, F. Shumin, L. Zhenning, Bus drivers' mood states and reaction abilities at high temperatures, Transport. Res. F Traffic Psychol. Behav. 59 (2018) 436–444, https://doi.org/10.1016/j.trf.2018.09.022.

[62] G. Yannis, E. Papadimitriou, K. Folla, Effect of GDP changes on road traffic fatalities, Saf. Sci. 63 (2014) 42–49, https://doi.org/10.1016/j.ssci.2013.10.017.

[63] T.Y. Chen, R.C. Jou, Using HLM to investigate the relationship between traffic accident risk of private vehicles and public transportation, Transport. Res. Pol. Pract. 119 (2019) 148–161, https://doi.org/10.1016/j.tra.2018.11.005.