Geographical spatial analysis and risk prediction based on machine learning for maritime traffic accidents

A case study of Fujian sea area

Yang, Yang; Shao, Zheping; Hu, Yu; Mei, Qiang; Pan, Jiacai; Song, Rongxin; Wang, Peng

DOI
10.1016/j.oceaneng.2022.113106

Publication date
2022

Document Version
Final published version

Published in
Ocean Engineering

Ocean Engineering 266 (2022) 113106

Contents lists available at ScienceDirect

Ocean Engineering

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

Geographical spatial analysis and risk prediction based on machine learning for maritime traffic accidents: A case study of Fujian sea area

Yang Yang [a,b], Zheping Shao [a], Yu Hu [b], Qiang Mei [a,c,*], Jiacai Pan [a], Rongxin Song [d], Peng Wang [e,c,**]

[a] Navigation Institute, Jimei University, Xiamen, 361021, China
[b] Xiamen Data Intelligence Academy of CAS, ICT, Xiamen, 361021, China
[c] Merchant Marine College, Shanghai Maritime University, Shanghai, 201306, China
[d] Safety and Security Science Group, Faculty of Technology, Policy and Management, Delft University of Technology, Delft, 2628 BX, the Netherlands
[e] The Institute of Computing Technology of the Chinese Academy of Sciences, Beijing, 100190, China

ARTICLE INFO

Keywords:
Geographical spatial analysis
Maritime accident
Fujian sea area
Machine learning
Accident prediction

ABSTRACT

Safety analysis according to the spatial distribution characteristics of maritime traffic accidents is critical to maritime traffic safety management. An accident analysis framework based on the geographic information system (GIS) is proposed to characterize the spatial distribution of maritime traffic accidents occurring in the Fujian sea area in 2007–2020 by employing kernel density estimation and spatial autocorrelation techniques. The sea area is divided into various grids, and in each grid, the mapping relationships between the number and severity of the traffic accidents and the traffic characteristics are established. Machine learning (ML) technology is used to assess whether a grid area is an accident-prone area and to predict accident severity in each grid. The accident prediction of different ML models, including random forest (RF) model, Adaboost model, gradient boosting decision tree (GBDT) model, and Stacking combined model, were compared. The optimality of the Stacking combined model was verified by comparing the experimental results of this model with those of classical prediction models, convolutional neural network (CNN), long short term memory (LSTM), and support vector machine (SVM). According to the results, the maritime accident data set of the entire Fujian sea area shows typical clustering characteristics and positive spatial correlation. That is, the kernel density estimation indicates that subareas, including the Ningde sea area, Fuzhou sea area, and Xiamen sea area, generally have high densities of maritime accidents and the highest risk value within the whole Fujian sea area. High-high accident clustering, that is high cluster areas neighbored by other areas of high cluster, is mainly seen in the Ningde and Fuzhou sea areas, while the Xiamen, Putian, and Zhangzhou subareas show low-low clustering, which are low clusters neighbored by low clusters. Among the ML models, the Stacking combined model shows high accuracy, precision, recall, and F1-score values of 0.912, 0.910, 0.912, and 0.904 in predicting whether a grid area is an accident-prone area and 0.750, 0.745, 0.750, and 0.746 in predicting the accident severity in the grid, indicating its superior maritime traffic accident prediction performance. Based on our analysis of the distribution characteristics and geospatial data, our proposed method demonstrates effective and reliable risk prediction.

1. Introduction

The size and speed of ships have greatly increased with the development of science and technology, along with economic globalization. Flourishing trade increases cargo transport volume, ship density in sea areas, and the intersection of shipping routes, making sailing conditions more complex. This has resulted in higher maritime accident risks (Dulebenets, 2018). A maritime accident is defined as an event directly resulting from the operations of a ship, and it causes any of the following consequences: death or serious injury; person missing from a ship; loss, presumed loss, or abandonment of a ship; material damage to a ship; stranding or disabling of a ship; collision; material damage to marine

* Corresponding author. Navigation Institute, Jimei University, Xiamen, 361021, China.
** Corresponding author. The Institute of Computing Technology of the Chinese Academy of Sciences, Beijing, 100190, China.
E-mail addresses: meiqiang@jmu.edu.cn (Q. Mei), wangp@ict.ac.cn (P. Wang).

https://doi.org/10.1016/j.oceaneng.2022.113106
Received 1 September 2022; Received in revised form 21 October 2022; Accepted 5 November 2022
Available online 24 November 2022

infrastructure external to a ship that could seriously endanger the safety of the ship, other ships or an individual; severe damage or the potential for severe damage to the environment brought about by the damage of a ship or ships (Maritime Safety Administration of People's Republic of China, 2010).

Maritime accidents seldom happen in normal scenarios but have catastrophic outcomes. Once accidents happen, people could suffer from huge losses of life and property, and the ecological environment may be irreversibly damaged (Yan, 2020). For example, the container ship EVER GIVEN running aground in the Suez Canal blocked the global supply chains for nearly a week in 2021, seriously impacting the shipping economy worldwide (Luo et al., 2022). It was estimated by the American Broadcasting Company that Egypt's daily economic loss was as high as $15 million, and the global trade volume decreased by $9 billion per day (www.cctv.com). Ships could bring a higher risk of accidents at sea if they carry dangerous goods, such as oil tankers (Nermin et al., 2022), because the entire sea area and the marine life there will be greatly affected once an accident happens.

Our research proposes a spatial analysis framework based on geographical characteristics for the spatial distribution of accidents in the Fujian sea for securing maritime safety. We identified the correlation between trajectory characteristics and the distribution of accidents. Furthermore, the geographic patterns of accidental events were determined and the geographic characteristics were discussed to predict trajectories more accurately. This paper focuses on two methods for reducing maritime accidents. First, the spatial distribution characteristics of accidents were analyzed using geographic information system (GIS) technology. Longitude, latitude, severity, and other data were extracted to determine the geographical location of the accidents and the distribution of accidents with different severities. The distribution of accidents and accident-prone areas were found using kernel density estimation. Then, the spatial autocorrelation method was used to determine the clustering scenarios of accidents with such a density distribution and to further obtain the local distribution characteristics of the accidents and specific accident clusters. Second, the characteristics of ship traffic flow were extracted from automatic identification system (AIS) data to predict accidents that have not yet occurred (Shu et al., 2017, 2018). As accident prediction is a basis for making a scientific safety decision, it is essential for accident prevention. Accident prediction is usually conducted by predicting future safety conditions of a system based on the past and present safety information of the system through a series of scientific methods (Guo et al., 2022). The spatial distribution characteristics of maritime accidents and accident prediction results can provide maritime authorities a more intuitive understanding of the traffic safety conditions of ships within their jurisdiction so that they can take targeted measures to reduce maritime accidents and ensure navigation safety (Wang et al., 2022).

2. Literature review

Characterization of the spatial distribution of accidents by using GIS technology is widely seen in many fields. Ma et al. (Ma et al., 2021a,b) used the density analysis method to identify the areas with high accident incidence and high accident severity based on the road traffic data of Wales in 2017. Then, they used two types of spatial clustering analysis models—outlier analysis and hot spot analysis—to further identify the regions with high accident severity and form the spatiotemporal distribution of accidents. Zhang et al. (2021) collected maritime accident data of 2003–2018 from the Marine Casualties and Incidents (MCI) module of the Global Integrated Shipping Information System (GISIS). Kernel density estimation and the K-means clustering method were used and manipulated, and descriptive analyses were carried out to obtain an overview of global maritime accidents. They found distributions of maritime accidents by time, initial event, and ship type were diverse in different accident classes. Yang et al. (2021) analyzed the spatial characteristics of maritime accidents occurring under specific meteorological conditions. After converting meteorological and environmental data and maritime accident data into spatial units, they clustered units with similar meteorological environmental conditions and compared maritime accident characteristics in each cluster. Wang et al. (2022) determined the spatial patterns of maritime accidents in terms of accident frequency and severity using the global maritime accident data from 2010 to 2019 by means of density analysis and clustering analysis. Their study could guide the relevant maritime authorities to improve maritime traffic management. Hammami and Matisziw (2021) suggested the possible existence of spatial and/or temporal dependencies (i.e., clusters or hot spots) among accidents.

Along with understanding where and when such spatiotemporal dependencies might occur, another important facet to consider is the geographic extent or area associated with the hot spots. Better delineation and quantification of the morphological characteristics of accident hot spots could provide valuable decision support for planning for accident hot spot mitigation and prevention. Misuk et al. (2021) suggested that risk factors threatening public safety, such as crime, fire, and traffic accidents, had spatial characteristics. Based on Global Moran's I, Local Moran's I, and Getis-Ord's G*I methods, they analyzed the spatial distribution pattern of the local safety level index and risk factors for each sector. Kalantari et al. (2021) proposed an exploratory spatial analysis framework for identifying and ranking hazardous locations of traffic accidents in Zanjan, one of the most populous and densely populated cities in Iran. This framework quantified the spatiotemporal association among collisions by comparing the results of different approaches, including the kernel density estimation, natural breaks classification (NBC), and Knox test. Feizizadeh et al. (2022) investigated the spatiotemporal trends of urban traffic accident hot spots during the COVID-19 pandemic. The severity index was used to determine high-risk areas, and the kernel density estimation method was used to identify the risk of traffic accident hot spots. This method identified the hot spots of urban traffic accidents and evaluated their spatiotemporal correlation with land use and demographic characteristics. Rong et al. (2021) presented a spatial correlation analysis method for near-collision clusters with local traffic characteristics. The Moran's I and Getis-Ord Gi* spatial autocorrelation methods were used to determine whether near collisions showed spatial clustering from global and local perspectives. Crimmins et al. (2021) suggested that many studies on traffic crashes considered various geometric roadway features; however, ever-evolving urban watersheds and climate change increasingly impacted roadway conditions. They used kernel density surfaces and local Getis-Ord Gi* statistics to identify locations prone to witnessing crashes in wet conditions. Local environmental and traffic risk factors were considered for the network performance evaluation. Katanalp and Ezgi (2021) conducted micro- and macro-level evaluations of pedestrian-vehicle crashes. Macro-level findings were obtained with GIS-based density analyzes, and critical road segments were determined. They established a converted fuzzy-decision model and a revised fuzzy-decision model. The results revealed that land use, parking, and peak hour volume greatly affected pedestrian safety, and the effects of public transport, speed, and road type were the greatest.

In order to recover or predict the maritime information and best mine vessels' data, the machine learning methods are applied to research (Liang et al., 2021; Yuan et al., 2020). Commonly used methods for accident prediction at present include Regression Forecast, Scenario Analysis, Time Forecast, Markov Chain Forecast, Gray Model, and Artificial Neural Networks. Lin and Li (2020) designed a hierarchical scheme for sequential prediction by using User-Generated Crowdsourcing Data (UGCD) for real-time Traffic Accidents Post-Impact (TAPI) prediction. The proposed model was validated by embedding three machine learning (ML) algorithms, random forest (RF), support vector machine (SVM), and neural network (NN). When the assessment was conducted under absolute difference conditions, the performances of the three models were ranked as follows: NN, RF, and SVM. Li et al. (2020) suggested that the fusion of features was an important factor in

predicting the duration of traffic accidents. They proposed a deep fusion model which could simultaneously handle categorical and continuous variables. In this model, a stacked restricted Boltzmann machine (RBM) was used to handle the categorical variables, a stacked Gaussian-Bernoulli RBM was used to handle the continuous variables, and a joint layer was used to fuse the extracted features. The proposed model could fully mine nonlinear and complex patterns in traffic accident data and traffic flow data. Chai et al. (2020) aimed to provide an efficient way to predict the number of vessel accidents in China. To weaken the randomness of the vessel accident number time series, the gray processing operation was adopted to generate a new sequence with exponential and approximate exponential rules. In addition, an extended least-squares support vector machine (LSSVM) model was applied in the forecasting of the new sequence. The parameters of the LSSVM were optimized by an improved quantum-behaved particle swarm (IQPSO). The proposed method proved to be effective in forecasting the number of vessel accidents in China.

The lockdown during COVID-19 has resulted in a lack of data on highway accidents involving the transportation of dangerous goods, thus affecting related research. Li et al. (2021) established the time series of accidents and an autoregressive moving average (ARMA) prediction model. The results indicated that the mean absolute percentage error (MAPE) between the actual and predicted values of transportation accidents was 0.147, 0.315, and 0.29. Therefore, the model met the prediction accuracy requirements. In the study of Yuan et al. (2021), the empirical probabilities of scenario nodes were obtained through defuzzification calculation, and the state probability of each scenario node was calculated by using the dynamic Bayesian network joint probability formula. A consequence prediction model was then established by constructing the correlation between the optimized scenario evolution path and the accident consequences. The occurrence probability of accident consequences was calculated by using the defuzzification method and dynamic Bayesian network. Ali et al. (2021) analyzed and predicted road traffic accidents (RTAs) using artificial neural networks (ANNs). Their model using the sigmoid activation function and Levenberg-Marquardt algorithm outperformed multivariate regression models. The model results indicated the estimated traffic accidents based on appropriate data were close enough to the actual ones. Xiong et al. (2021) summarized the influencing factors of freeway traffic safety as human behavior characteristics, vehicle factors, road factors, environmental factors, and traffic safety factors after a systematic analysis. They measured the freeway safety level by using the hierarchical entropy method and predicted future traffic accidents in the sample area by using the autoregressive integrated moving average (ARIMA) model. The average error rate of prediction was only 0.47%, showing a high degree of fitting and accuracy. In the study of Kim et al. (2021), several ML models were applied to predict accidents at a container port under various time intervals. The optimal model was selected by comparing the accuracy, precision, recall, and F1 score of different models. The deep neural network model and the gradient boosting model exhibited the highest performance in terms of all the performance metrics. The applied methods could be used in predicting accidents at container ports in the future. Kumar et al. (2022) used classification models, specifically logistic regression, artificial neural network, decision Tree, K-nearest neighbors, and random forest, to predict the accident severity. Their study aimed to determine the specific features which could affect vehicle accident severity. The decision tree model was found to be the optimal model.

To update the current highway design criteria, Macwdo et al. (2021) proposed an accident prediction model for rural roads with single lanes by using a geographic information system. The geometric reconstruction of the original vector data and the semi-automatic extraction of the target road sections from the satellite images were conducted with the least statistically significant variables. The homogeneous segments were analyzed and classified by using the spatial method (kernel-KDE density). The generalized estimation equation (GEE) model was used to model the frequency and severity of accidents. The results revealed that expanding slope and radius could increase the frequency of curve accidents but reduce their severity. By analyzing the contributing factors that affect injury severity to facilitate the prediction of injury severity, Ma et al. (2021) developed an effective stacked sparse autoencoder (SSAE)-based analytic framework to predict the severity of traffic accident injuries. Based on geographical information, he classified the data using an SSAE-based deep learning model to efficiently predict injury severity. Yang et al. (2022) proposed the deep neural network (DNN) model to accurately predict traffic accident severity risks based on Chinese traffic accident data. This paper discusses a multi-task DNN framework constructed to predict different levels of injury, death, and severity of property loss in traffic scenarios.

This case study proposes a prediction framework of "AIS data + GIS preprocessing + ML prediction" by combining GIS and ensemble ML algorithms to predict and analyze "whether a grid area is an accident-prone area" and "the accident severity". The Fujian sea area is taken in this case study. The data on maritime accidents there from 2007 to 2020 are utilized. The training effects of different ML models–RF model, Adaboost model, gradient boosting decision tree (GBDT) model, and Stacking combined model–are compared. Then the optimal model is selected to predict accidents.

3. Methodology

Maritime accident data from 2008 to 2020 published by Fujian Maritime Safety Administration of China are used for this study. Each record includes information on the latitude and longitude, time, type, and severity level of the accidents. The original data were preprocessed and imported into ArcGIS for spatial pattern analysis (Ye et al., 2022). Kernel density analysis was conducted to determine the density distribution of these accidents and accident-prone areas. The spatial autocorrelation method was then used to analyze the clustering of accidents with such a density distribution. Next, the local distribution characteristics and specific accident clusters were obtained. The above results are used as inputs to different ML models.

The sea area is divided into standardized statistical grids. The traffic flow characteristics of different dimensions in each grid area are used as inputs for ML to predict accidents. The training and prediction effects of different ML models are evaluated and compared to investigate the feasibility of the accident prediction ML model. The structure of the framework is illustrated in Fig. 1.

3.1. Density analysis

Kernel density estimation is used for cluster analysis in this paper. The accident spots and their surrounding regions are considered to reflect the accident distribution in this area. Density analysis can process known values of a phenomenon and spread them across the landscape

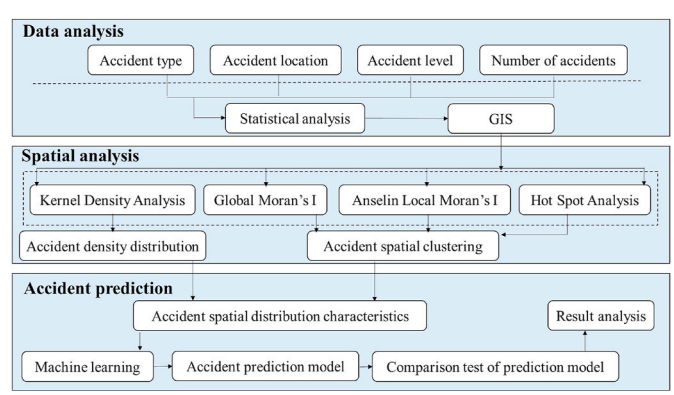

Fig. 1. The structure of the proposed framework.

according to the spatial relationship between the values measured in each spot and the locations of each value. The kernel density analysis calculates the density of point features around each output raster cell. It can calculate the density of both point and line features.

In this study, the studied area is divided into various grids. Conceptually, each point is covered by a smoothly curved surface. The surface value is the highest at the location of the point and decreases with the increase of the distance from the point, reaching zero at the search radius distance from the point. Only a circular neighborhood is allowed. The volume under the surface equals the Population field value for that point, or 1 if the field value is specified as NONE. The density at each output raster cell is calculated by adding the values of all the kernel surfaces where they overlay the raster cell center. *Density* is the predicted density of the accident spot, *i* is the accident spot, *popi* is an optional parameter, *disti* is the distance between accident spot *i* and other locations of accident spots, and *radius* is the default search radius. The following formulas (Silverman, 1998) define how to calculate the kernel density for a point and how to determine the default search radius within the kernel density formula.

$$Density = \frac{1}{(radius)^2}\sum_{i=1}^{n}\left[\frac{3}{\pi}\bullet popi\left(1-\left(\frac{dist_i}{radius}\right)^2\right)^2\right],\ for\ dist_i < radius \tag{1}$$

$$SearchRadius = 0.9 * \min\left(SD, \sqrt{\frac{1}{in(2)} * D_m}\right) * n^{-0.2} \tag{2}$$

The search radius refers to the data range involved in the kernel function. A larger search radius means a larger range of data, which makes the result more abstract. A smaller search radius contains more details, which makes the results more fragmented. *Dm* is the (weighted) median distance from the (weighted) mean center. *n* is the number of points if no population field is used. If a population field is supplied, *n* is the sum of the population field values. SD is the standard distance. *min* means that whichever of the two options, either SD or $\sqrt{\frac{1}{in(2)} * D_m}$, that results in a smaller value will be used. There are two ways to calculate the standard distance, unweighted and weighted.

The unweighted distance formula can be expressed as:

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - X^2}{n} + \frac{\sum_{i=1}^{n}(y_i - Y^2}{n} + \frac{\sum_{i=1}^{n}(z_i - Z^2}{n}} \tag{3}$$

where *xi*, *yi*, and *zi* are the coordinates for feature *i* (accident spot); $\{\overline{X}, \overline{Y}, \overline{Z}\}$ represents the mean center for the features; *n* is the total number of accident spots.

The weighted distance formula is:

$$SD_w = \sqrt{\frac{\sum_{i=1}^{n} w_i(x_i - \overline{X}_w)^2}{\sum_{i=1}^{n} w_i} + \frac{\sum_{i=1}^{n} w_i(y_i - \overline{Y}_w)^2}{\sum_{i=1}^{n} w_i} + \frac{\sum_{i=1}^{n} w_i(z_i - \overline{Z}_w)^2}{\sum_{i=1}^{n} w_i}} \tag{4}$$

where w_i is the weight of feature *i* and {x w, y w, z w} is the weighted mean center.

3.2. Spatial auto-correlation analysis

The spatial autocorrelation coefficient is often used to quantitatively describe the spatial dependence of accident spots. In this paper, Moran's I method is used to explore the spatial autocorrelation between accident spots, which reflects the spatial cluster of these spots. In the global Moran's analysis, significant Moran's I suggests a spatial correlation between accident spots in the area. However, in most cases, Local Moran's I (Local Moran index) is needed for an additional explanation because the specific accident location in the spatial cluster is not clear. Hot spot analysis is used as a supplementary tool for Local Moran's I to analyze accident hot spots, supplementing and verifying the spatial distribution of the accident dataset.

The Spatial Statistics toolbox contains statistical tools for investigating spatial distributions, patterns, processes, and relationships. In terms of concept and objective, there may be some similarities between spatial statistics and non-spatial statistics (using traditional methods). However, spatial statistics are unique since they have been developed specifically to deal with geographic data. Unlike traditional non-spatial statistical analysis methods, spatial statistics incorporate geographic space (proximity, area, connectivity, and/or other spatial relationships) directly into mathematics. Spatial statistics tools can be used to summarize salient accident characteristics (for example, determine the mean center or overall direction trend), identify statistically significant spatial clusters for accidents (hot/cold spots) or spatial outliers, assess overall accident patterns of clustered or dispersed, group accidents according to attribute similarities, determine the appropriate analysis scale, and explore spatial relationships.

3.2.1. Global Moran's I

Global Moran's I tool in the Spatial Statistics toolbox is used to measure the spatial autocorrelation based on feature locations and attributes. The accident dataset is imported into the tool, to evaluate whether the pattern expressed is clustered, dispersed, or random. The tool calculates the Moran's I index value, z-score, and p-value to evaluate the significance of the index. The P-value is the approximation of the area under the curve with a known distribution (limited by the test statistics). The formulas are presented as follows (Getis and Ord, 2010):

$$I = \frac{n}{s_0}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}z_i z_j}{\sum_{i=1}^{n} z_i^2} \tag{5}$$

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} \tag{6}$$

where $w_{i,j}$ is the spatial weight between feature *i* and *j*. z_i and z_j are the normalized observed values of the accident spots in space units *i* and *j*. *n* is equal to the total number of accident spots. S_0 is the aggregate of all spatial weights.

The Z_I-score for the statistics is computed as:

$$Z_I = \frac{I - E[I]}{\sqrt{V[I]}} \tag{7}$$

where:

$$E[I] = -\frac{1}{(n-1)} \tag{8}$$

$$V[I] = E\left[I^2\right] - E[I]^2 \tag{9}$$

where *V[I]* and *E*[*I*] are the variance and expected value of Moran's I.

3.2.2. Anselin Local Moran's I

Anselin Local Moran's I tool can be used to identify statistically significant spatial accident clusters with high or low values. The tool can be used to determine the sharpest boundaries of accident-prone areas and areas with higher accident severity and to determine an unusual accident spot. It can also identify whether there are unexpected accident spots, and the accident locations of unexpectedly high rates. The formulas are given as (Luc Anselin, 1995):

$$I_i = \frac{x_i - \overline{X}}{S_i^2}\sum_{j=1, j\neq i}^{n} w_{i,j}\left(x_j - \overline{X}\right) \tag{10}$$

$$S_i^2 = \frac{\sum_{j=1, j\neq i}^{n} \left(x_j - \overline{X}\right)^2}{n-1} \tag{11}$$

where x_i is the accident level of feature i; $\overline{X}$ is the mean of the corresponding attribute; $w_{i,j}$ is the spatial weight between features i and j; S_i is the aggregate of all spatial weights. n is equal to the total number of accident spots.

The z_{I_i}-score for the statistics are calculated as:

$$Z_{I_i} = \frac{I_i - E[I_i]}{\sqrt{V[I_i]}} \tag{12}$$

where:

$$E[I_i] = -\frac{\sum_{j=1, j\neq i}^{n} w_{i,j}}{n-1} \tag{13}$$

$$V[I_i] = E\left[I_i^2\right] - E[I_i]^2 \tag{14}$$

where $E[I_i]$ and $V[I_i]$ are the expected value and variance value.

3.2.3. Hot spot analysis

Accident severity is selected as the field. The Getis-Ord Gi* tool can be used to determine the location of spatial accident clusters with high or low values. The Getis-Ord statistics are used in hot spot analysis (Yang et al., 2022) to determine whether a point belongs to the same category as its neighbors (Getis and Ord, 2010). A high value of the Getis-Ord statistic indicates a cluster of high index values (hot spots), while a low value indicates a cluster of low index values (cold spots). The G_i^* statistic of Getis-Ord can be determined using the following equation:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j} x_j - \overline{X \sum_{j=1}^{n} w_{i,j}}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}} \tag{15}$$

The G*i statistic obtained is a z-score. For statistically significant z-scores (z-score>1.96 or < –1.96), the larger the z-score is, the more intense the clustering of high values (hot spot) would be. Conversely, the smaller the z-score is, the more intense the clustering of low values (cold spot) (Ord and Getis, 1995) would be.

3.3. Machine learning models

Chen et al. established the integrated learning yield prediction model to predict the yield of a fruit tree, which had the best prediction performance compared with traditional prediction models, such as the support vector regression (SVR) model and K-nearest neighbor (KNN) model (Chen et al., 2022). Bai et al. used the integrated learning model to predict the thermal comfort of building occupants and showed through systematic comparison that the prediction performance of the integrated learning model outperformed 10 other machine learning models trained with different data subsets (Bai et al., 2022). Shan et al. proposed a prediction model based on integrated learning to predict hourly solar irradiance, compared it with traditional prediction models, such as the RF and SVR models, and showed that the integrated learning model had the highest prediction accuracy (Shan et al., 2022). Li and Song (2022) used the integrated learning model to predict the strength of high-performance concrete, which again showed better performance compared with other ML models.

Based on literature review, this study used the integrated learning algorithm to assess whether a grid area is an accident-prone area and to predict accident severity in each grid. Moreover, we compared different integrated learning models, including the RF model, Adaboost model, GBDT model, and Stacking combined model. Lastly, the best-performing model was selected to predict and analyze accidents, as well as demonstrate correlations between the AIS data processed by GIS and the accident. The integrated learning model is shown in Fig. 2.

Different ML models (i.e., RF model, Adaboost model, GBDT model and Stacking combined model) are applied to identify "whether a grid area is an accident-prone area" and determine "the accident severity". After the training performance being evaluated and compared, an optimal model is selected for accident prediction based on the correlation between the accidents and the AIS data processed by GIS.

Each ML model ensembles a set of learning algorithms. High-level classification performance was achieved by constructing a strong learner from a combination of several weak learners using the following two methods:

① Bagging：

Bagging is known as bootstrap aggregation. It can be used for reducing variance within a noisy dataset. In bagging, a new sample is established to represent the distribution of the original sample through resampling from a limited number of samples (uniform sampling from a given training set data with replacement, that is, every time a sample is selected, it is equally likely to be selected again and re-added to the training set).

Bagging algorithm:

Several independent learners are constructed using a base learner algorithm on the basis of bootstrap dataset. These learners are called base learners (also known as homogeneous learners) and are commonly generated by using the same base learning algorithm. Individual learners can be generated in parallel without strong interdependence. The classification results of these base learners are summarized or averaged through the majority voting mechanism as the final result.

The Random Forest model is representative of Bagging construction and is used to analyze maritime accident data in this paper.

② Boosting:

Boosting methods give repeat training to the data and allocating different weights to the data to get a number of weak classifiers (also known as basic classifiers). These weak classifiers are combined to form a strong classifier. The whole process can be divided into two stages: continuous repeated learning and a combination of different learners.

In the first stage, most Boosting methods change the probability distribution of training set data (weights of different samples of training data) and call the weak classifier algorithm according to the data with different probability distributions.

Typical Boosting-related models include Adaboost and GBDT, and they are used for the analysis of maritime accident data in this paper.

3.3.1. Random forest

The RF model is an integrated machine learning algorithm constructed from decision trees. The model has a clear structure, is easy to explain, has high stability, and is not prone to overfitting. It is commonly used for the classification of applications (Pei et al., 2022).

RF uses the bootstrapping method. In the given m sample datasets, after n times of random sampling, n training sample sets are obtained. Each sample set is trained to construct decision trees. At the node of the decision tree, a subset of k attributes is randomly selected from the node attribute set, among which an optimal attribute is selected for splitting. Test samples will be input into each decision tree for classification output or regression output after the establishment of the random tree. For classification, the final result is determined by voting (Xing et al., 2021).

Using averages or majority voting, the probability of judgment error

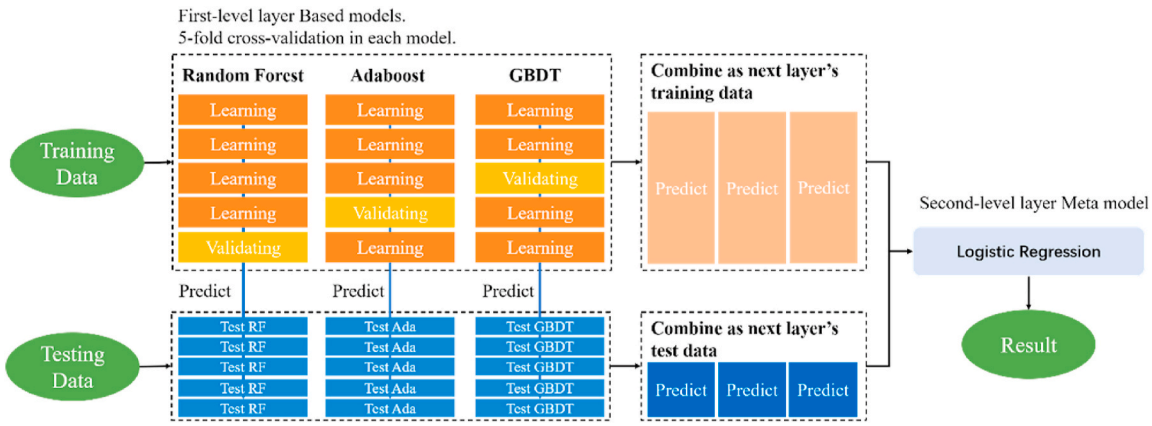

Fig. 2. Accident prediction process by combined models.

(e_rfc) for any accident sample can be written as:

$$e_rfc = \sum_{i=n}^{N} C_N^i \varepsilon^i (1-\varepsilon)^{N-i} \tag{16}$$

where i represents the number of judgment error; ε is the probability of wrong judgment of a tree while $(1-\varepsilon)$ is the probability of correct judgment; $(N-i)$ is the total correct judgment times.

3.3.2. Adaboost

Adaboost algorithm is a multi-learner enhancement technique based on ML. A base learner is an algorithm set composed of one or more ML algorithms, and it can be used as a basic unit to judge the maritime accident level. The Adaboost algorithm can integrate multiple weak base learners (base learners with low accuracy in identifying accident types). Weak base learners are combined to form a strong base learner by adjusting the weights to the weak base learners. Thereby, the classification accuracy of the Adaboost algorithm is improved (Li et al., 2022).

The linear combination of Adaboost algorithm base learners and the exponential loss function of these base learners are expressed as:

$$H(x_i) = \sum_{t=1}^{T} \alpha_t h_t(x_i) \tag{17}$$

$$L_{exp}[H(x_i|D)] = E_{xi.\sim D}\left[e^{-f(x_i)H(x_i)}\right] \tag{18}$$

where $h_t(x_i)$ represents the base learners; α_t is the weight coefficient of the base learners. T is the number of base learners; $f(x_i)$ is the classification of parameter x_i. D is the parameter distribution. The sign function is introduced to minimize the exponential loss function, which can be expressed as:

$$G(x) = sign(H(x_i)) = sign\left(\left(\sum_{m=1}^{M} \alpha_t h_t(x_i)\right)\right) \tag{19}$$

where $G(x)$ is the final strong classifier. If the value is greater than 0, the output of the strong classifier will be 1. If the value is less than 0, the output will be −1. If the value is 0, the output will be 0.

3.3.3. Gradient boosting machine

The GBDT model is an emerging ML method that classifies data using an additive model (a linear combination of decision trees as basic functions) and continuously reduces the residuals generated during trainings. It is a Boosting algorithm. Based on the previously built model loss function, each time a smaller loss function is taken along the gradient descent direction. Thus, an improved learner is established. Larger loss function means more mistakes made by the model. Continuous decrease of the loss function means an improving model it is becoming. The best way to improve the model is to let the loss function decrease along its gradient direction. Theoretically, the gradient boosting machine can employ different learning algorithms as the base learner.

In GBDT, the decision Tree is used as the base learner for the gradient boosting machine. After a decision tree is constructed, the residual outputs of the existing model and the actual output from samples are used to construct another tree. Through successive iterations, the results of all decision trees are taken as the output (Shen et al., 2022). The formula is as follows:

$$f_m = \sum_{m=1}^{M} T(x;\theta_m)\Bigg) \tag{20}$$

where x is the characteristic variable; T stands for the decision tree; θ is the parameter of the decision tree; M is the number of trees.

Root mean square error (RMSE) is used as the evaluation index for the generalization ability of the model. The formula is written as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(f_{ir} - f_{ip}\right)^2} \tag{21}$$

where f_{ir} is the true value; f_{ip} is the predicted value; N is the number of accident training samples.

3.4. Deep learning and traditional machine learning algorithms used in this study

Convolutional neural networks (CNNs) are feedforward neural networks with convolutional computation and deep structure and are one of the representative algorithms of deep learning (Gu et al., 2018). CNNs have representation learning and shift-invariant abilities to classify input information according to their hierarchical structure. Therefore, they are also called "shift invariant artificial neural networks" (SIANN) (Artyomov and Yadid Pecht, 2005).

Long short term memory (LSTM) networks are cycle-time neural networks specially designed to solve long-term dependency problems of general recurrent neural network (RNN). All RNNs have chains of repetitive neural network modules ().

The support vector machine (SVM), a traditional machine learning method, is a two-group classification model. Its baseline model is the linear classifier with the largest interval defined in the feature space. SVM learns by maximizing the interval, which can be shown as a quadratic convex optimization programming problem. The learning algorithm of the support vector machine is the optimal algorithm for solving convex quadratic programming (Blanco et al., 2022).

3.5. Data description

3.5.1. Marine accident data of Fujian Province

The Fujian sea area in Fujian Province, China is selected as the studied area. Fujian, or "Min" for short, is located on the southeast coast of China. It borders Zhejiang Province in the northeast, Jiangxi Province in the west and northwest, Guangdong Province in the southwest, and Taiwan Province across the Taiwan Strait in the east. It is an important seaport in the Chinese mainland and a window for China to communicate with the world. It has the second longest coastline in China, with a zigzag coastline of 3751.5 km and a sea area of 136,300 square kilometers (Fig. 3).

3.5.2. Statistical analysis of research data

The accident data from January 2007 to April 2020 are from Fujian Maritime Safety Administration of China. They include a total of 549 records, mainly including accident location, time, type, severity level, and economic loss. The accidents are classified into "minor accident", "ordinary accident", "major accident" and "serious accident" by severity (MSA, 2010, Table 1). They can also be classified into "touch rocks", "stranding", "collision", "touch", "fire and explosion", "sank", "operational pollution", "damage by waves", "wind" and "others" according to the accident type (MSA, 2010, Table 2).

The severity of maritime traffic accidents is classified through such factors as casualties, direct economic losses, or environmental pollution of water areas. Minor accidents are accidents below the ordinary level. In this study, the accidents are classified based on casualties (Table 1). Minor accidents account for the highest proportion, followed by ordinary accidents. Major and serious accidents account for a relatively small proportion. However, their impacts, like ship accidents, casualties, economic losses, and environmental pollution, are far greater than those of minor and ordinary accidents. Therefore, the prevention of major and serious accidents should be the focus of attention. Table 2 lists ten types of maritime traffic accidents in the Fujian sea area, among which "collision" (damage caused by a collision between two or more ships) account for the highest proportion, followed by "touch rocks" and "touch". The occurrence frequency of accidents caused by "waves" or "wind" in this area is relatively low.

The total number and severity of accidents in cities like Fuzhou,

Table 1
The number of accidents of each severity level.

S/N	Accident severity	Number	Definition (casualties)
1	Minor accidents	324	No casualties
2	Ordinary accidents	124	Less than 1–3 people died (including missing)
3	Major accidents	75	Less than 3–10 people died (including missing)
4	Serious accidents	26	Death (including missing) of less than 10–30 people

Table 2
Accident type.

S/N	Accident type	Number	Frequency
1	Touch rocks	65	middle
2	Stranding	41	middle
3	Collision	249	high
4	Touch	69	middle
5	Fire and explosion	29	middle
6	Sank	56	middle
7	Operational pollution	3	low
8	Damage by waves	1	low
9	Wind	4	low
10	Others	32	middle

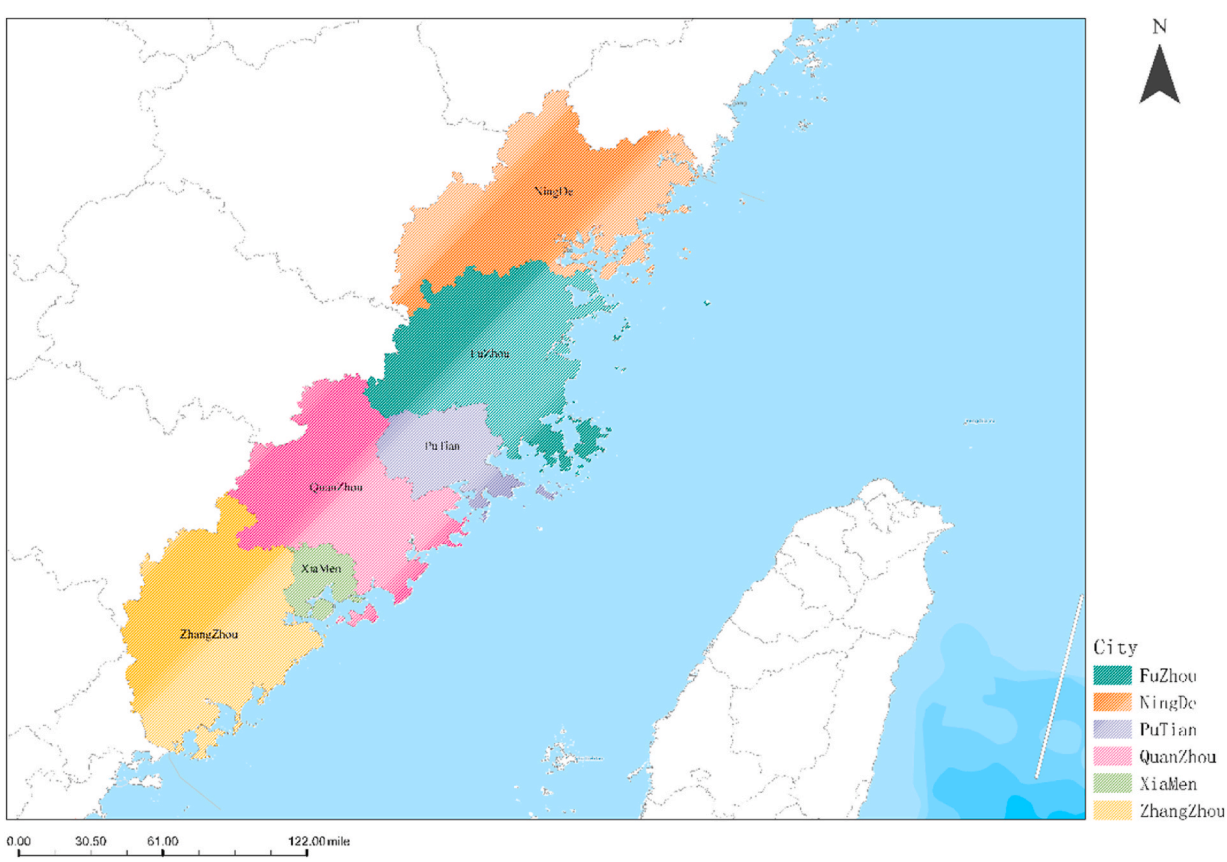

Fig. 3. Research area–Fujian Province, China.

Ningde, Putian, Quanzhou, Xiamen, and Zhangzhou are compared. Fuzhou witnessed the most accidents, followed by Xiamen, Zhangzhou, and Ningde. In Putian and Quanzhou, accidents are fewer. Fuzhou and Ningde had the highest proportion of serious accidents. Fuzhou, Ningde, and Xiamen have the highest proportion of major accidents. Minor accidents account for the highest proportion in all regions. Serious accidents rarely occur in Putian and Zhangzhou (Fig. 4).

3.5.3. The spatial distribution of AIS data

The AIS data of the Fujian sea area, a total of 8,340,427 items obtained from Shanghai Maritime University, were also used for presenting information about ship trajectories and the corresponding times. They were used to describe ship behavior and could be applied to fields like port traffic flow, port ship network analysis, and ship accident analysis. AIS data can reflect the flow and motion of ships effectively and directly. After the AIS data are input into GIS software, the framework for ship accident prediction is established by combining the maritime accident investigation data with other data—ship traffic flow, the average and standard lengths and widths of ships, course and speed standard deviations, and mean values. As shown in Fig. 5, different color blocks represent the density of traffic flow in different regions. The darker the color is, the higher the density would be. Red blocks mean the highest traffic density, followed by yellow ones and then green ones. Subareas like the Ningde sea area, Fuzhou sea area, Quanzhou sea area, and Xiamen sea area are regions with high traffic density, followed by the Putian sea area and Zhangzhou sea area. Fujian with a long coastline, vast sea area, and rich marine resources has huge potential for marine development and economic growth. The traffic density of this port is high in this prosperous sea area in the west of the Taiwan Strait.

4. Analysis and results

4.1. Accident spatial distribution

The visual spatial layout of maritime accident spot distribution and accident type distribution in the Fujian sea area is obtained by locating maritime traffic accidents in the map layer based on the longitude and latitude coordinates and the types of traffic accidents (Fig. 6). The green dots in Fig. 6 (A) represent the spatial distribution of all maritime traffic accidents in the Fujian sea area. In Fig. 6 (B), the purple dots represent grounding accident, the orange dots represent operational pollution accident, the blue dots represent wave damage accident, the yellow dots represent fire and explosion accident, the red dots represent collision accident, the green dots represent sinking accident, the brown dots represent touch accident, the light blue dots represent reef accident, the dark purple dots represent wind disaster, and the dark green dots represent other accident types.

4.2. Kernel density analysis

The number of traffic accidents in per unit channel is an important index for assessing traffic safety in navigation. The kernel density analysis method can be used to determine the spatial distribution of traffic accidents in the Fujian sea area and then the navigation section with high accident frequency can be identified. Based on the maritime traffic accident data in this sea area, the spatial distribution density of accident spots can be determined using Equation (1). The kernel density values of the elements in the whole analysis area are superimposed to calculate the kernel density values of all grids in the area (Fang et al., 2021).

The kernel density spatial distribution results calculated from maritime accidents in the Fujian sea area are shown in Fig. 7. Different color blocks represent different densities, and the chroma ranges from light to dark. The darker the chroma is, the higher the aggregation and the accident density can be. The accident set shows different clusters to spatial distribution and has different cluster centers (with the highest risk value). The main cluster centers locate in subareas including the Xiamen sea area, Ningde sea area, and Fuzhou sea area. The secondary cluster centers are distributed in subareas including the Zhangzhou sea area, Quanzhou sea area, and Putian sea area. The areas beyond cluster centers with chroma not so dark indicate that the accident rates are not high in these areas. Given the high cluster degree of main cluster centers, the situation of the Fujian sea area was analyzed. The situation of Xiamen Port is relatively complex: high-risk cluster centers are almost distributed all over the port. Xiamen Port consists of six port areas, and its main cluster centers are Xiamen Bay and the west sea area. There are three port areas (Houshi Port Area, Zhaoyin Port Area, and Haicang Port Area) and four wharves (Haitian Wharf, Youlun Wharf, Haitong Wharf, and Songyu Wharf) in the main channel of Xiamen Bay. The secondary cluster center of Xiamen Port is located near the east sea area, Xiang'an Port Area, Big Kinmen Island, Beidong Waterway, and Huyu Island. Results show that Sandu Island near Ningde Port is one of the cluster centers bearing the highest risk of accidents, and Dongan Island and Leijiang Island are the secondary cluster centers of this port. Fuzhou Port has two cluster centers with high value at risk; one is Minjiangkounei Port Area, and the other is Songxia Port Area. Dongshan Port Area is the high-risk cluster center of Zhangzhou Port. Quanzhou Bay Port Area is the high-risk cluster center of Quanzhou Port, and Xiuyu Port Area is the high-risk secondary cluster center of Putian Port.

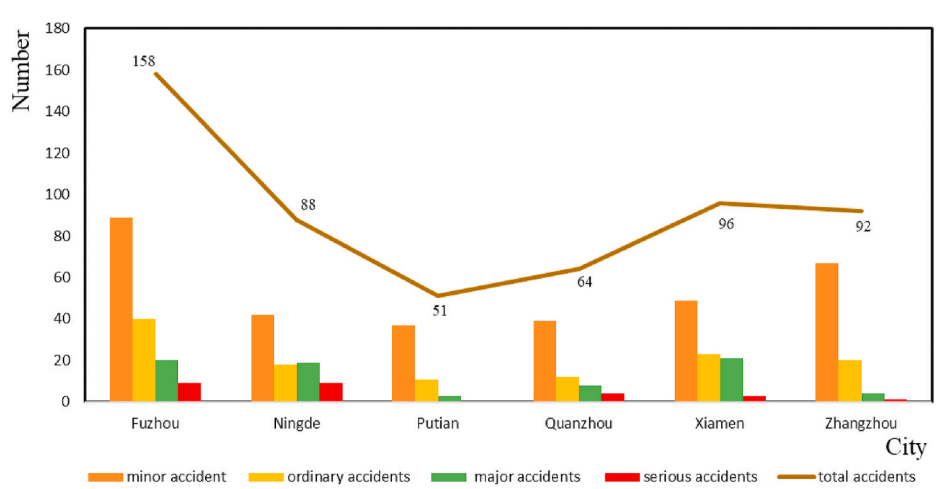

Fig. 4. Comparison of accident number and severity of each city.

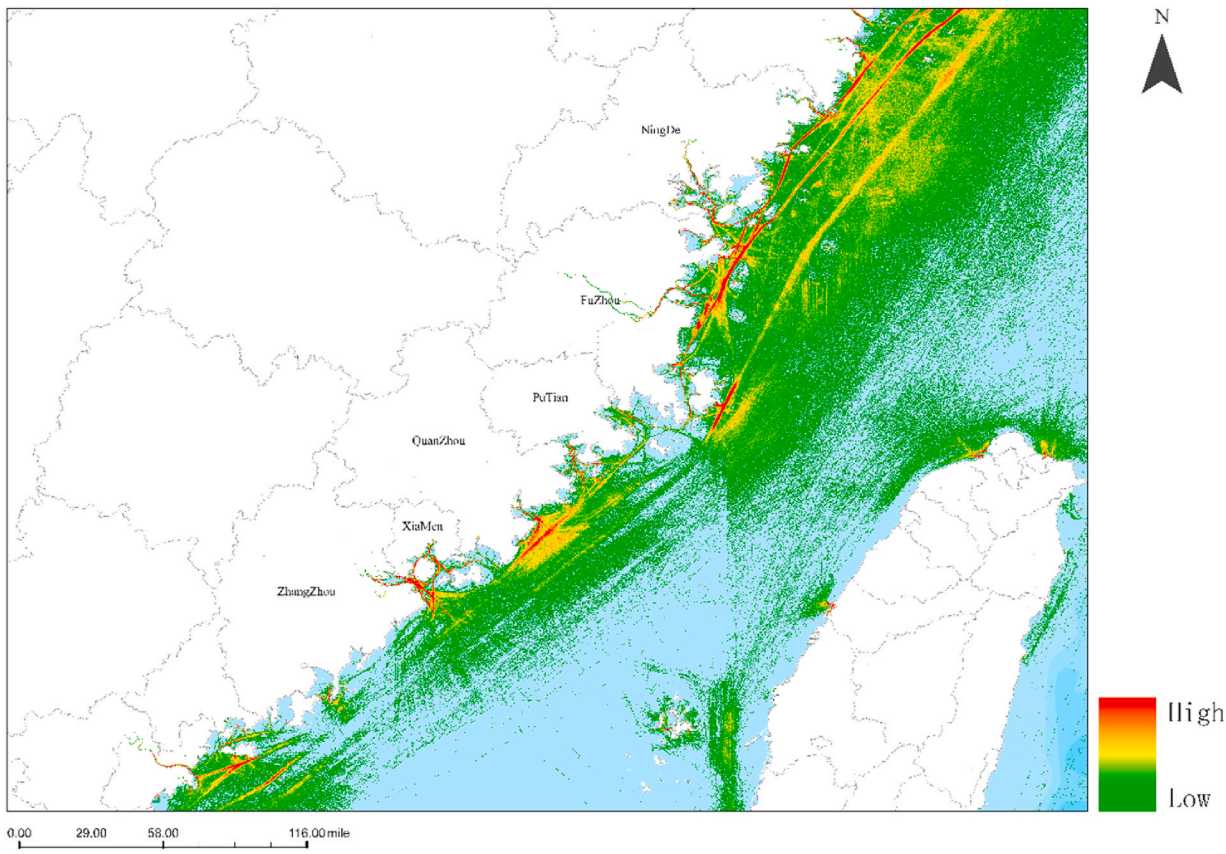

Fig. 5. Traffic flow in the Fujian sea area.

4.3. Spatial autocorrelation analysis

4.3.1. Global Moran's I analysis

Table 3 and Fig. 8 show the details of Global Moran's I analysis in the Fujian sea area. After the data are selected, the spatial autocorrelation tool in the GIS toolbar is applied: setting the characteristic field as accident severity, the distance threshold value 400. After the numerical value is input, the spatial autocorrelation tool will return five values: Moran's I index, expected index, variance, z-score, and p-value. In the Moran's index, p-values and z-scores are used to determine the spatial correlation, and the variance reflects the degree of dispersion between accident spots. P-values are probability values, and z-scores represent multiples of standard deviations. The confidence of spatial autocorrelation can be obtained by correlating p-values and z-scores with the Moran's index. The confidence interval refers to the estimation interval of the overall accident parameters constructed by the accident sample statistics. The greater the confidence interval is, the higher the occurrence probability of outliers can be. Thus, the occurrence probability of clustering or outliers in the accident set of the Fujian sea area can be determined. Local autocorrelation can be performed if global autocorrelation occurs, because local Moran's I can locate outliers and cluster points.

Fig. 8 shows that the Moran's I report presents a significant normal distribution and is divided into three parts. The middle part is the random distribution, the right part is clustered distribution, and the left part is dispersed distribution. In Table 3, its Moran's index is about 0.057, and its z-score is about 3.90, which is about 3.90 times the standard deviation. The value of Moran's index being positive indicates that the results that are distributed on the rightmost of the normal distribution show a clustered distribution. The p-value is about 0.000095, suggesting that the result is entirely not generated by random data and the result is reliable. As the probability of randomly generating this clustering pattern is less than 1%, it indicates that the data are clustered with 99% certainty and the probability of data clustering is greater than that of random distribution. Therefore, the spatial distribution of maritime traffic accidents in the Fujian sea area show some clustering characteristic, and it is a spatial positive correlation pattern.

4.3.2. Anselin Local Moran's I

Based on the spatial autocorrelation analysis results, the accident set is internally correlated and featured as clustered distribution, but the exact spatial cluster points remain unknown. Local Moran's I autocorrelation analysis is performed based on Moran's I analysis to determine the exact spatial cluster points (Zhang et al., 2019). The clustering results of traffic accident distribution in the Fujian sea area obtained by Local Moran's I autocorrelation analysis is presented in Fig. 9. The selected characteristic field is accident severity.

In Fig. 9, the gray points indicate not significant difference in accident severity; the pink points represent the high-high cluster, i.e., spots around these points have accidents with high severity; the red points indicate high outlier, i.e., spots with high accident severity are surrounded by those with low accident severity; the blue points represent the low-low cluster, i.e., all accidents happening around these points are of low severity; the dark blue points indicate low outlier, i.e., spots with low accident severity are surrounded by those with high accident severity. Spots with high accident severity are mainly located in the Ningde Port Area and Fuzhou Port Area. Larger number of accidents with high severity in these two ports may form the low outlier as the places with low accident severity are almost surrounded by those with high accident severity. While minor accidents and ordinary accidents are likely to occur in Putian Port, Xiamen Port, and Zhangzhou Port. Larger number of accidents with low accident severity in these three ports may form the high outlier points since the places with high accident severity are almost surrounded by those with low accident severity.

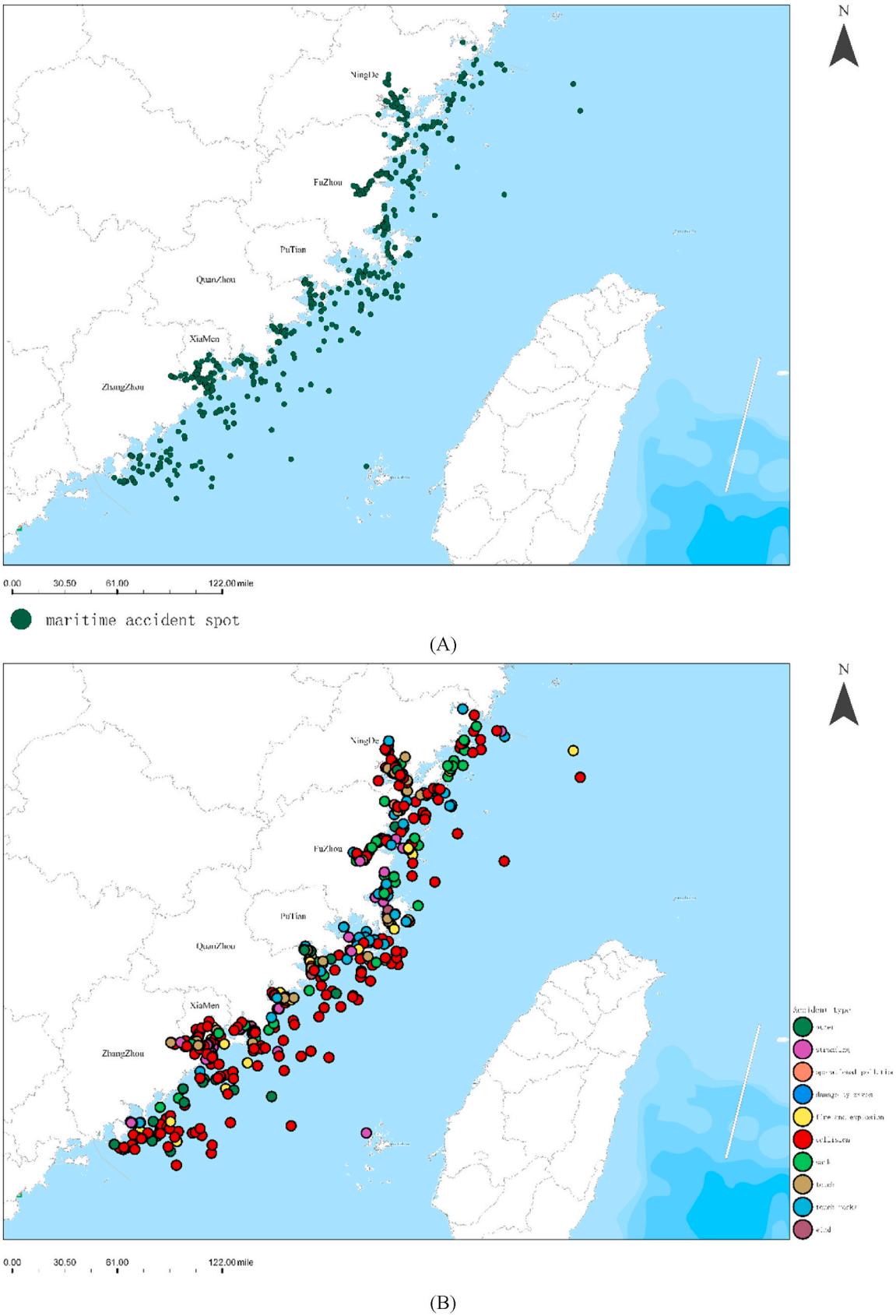

Fig. 6. The visual spatial layout of maritime accident spot distribution (A) and accident type distribution (B) in the Fujian sea area.

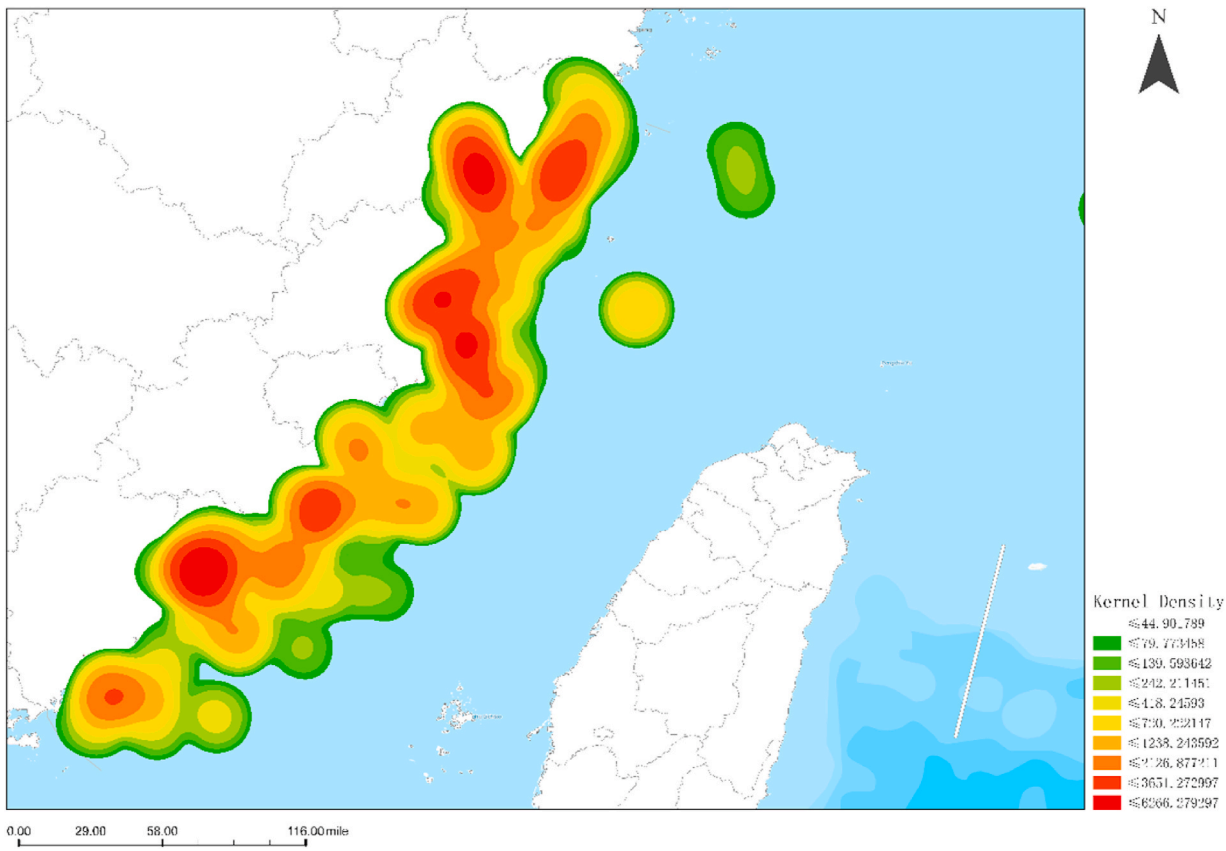

Fig. 7. Spatial distribution of maritime accident kernel density in the Fujian sea area.

Table 3
Spatial autocorrelation report.

Correlation value	Moran's I index	Expected index	Variance	z-score	p-value
Score	0.057867	−0.001825	0.000234	3.903731	0.000095

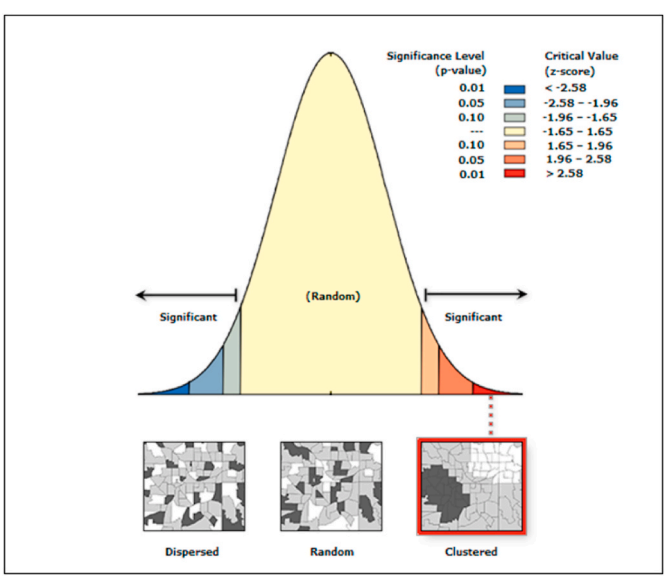

Fig. 8. Spatial autocorrelation report.

Gray points are distributed all along the Fujian sea area, indicating that the accident severity around the accident spot could not form high-high or low-low clusters.

4.3.3. Hot spot analysis

The hot spot analysis can be used to calculate the Getis-Ord Gi* statistics (called G-i-asterisk) for each accident in the accident set. The obtained z-score and p-value can determine the location of spatial clusters of high or low accident severity values. This analysis method works by checking each accident element in the immediate accident environment. High severity accident elements tend to attract attention but may not be statistically significant hot spots. Hot spot with statistical significance means that an accident of high severity should have a high value and be surrounded by other accidents of the same severity. An accident element and its local sum of accident elements are compared with the sum of all accident elements. When the local sum is different from the expected local sum making it fail to be randomly produced, a statistically significant z-score will be produced. Fig. 10 shows the hot spot analysis of the Fujian sea area.

In Fig. 10, the croci, orange, and red points are called "hot spots" and represent sites where major accident levels cluster; the inky blue, light blue, and light gray points are called "cold spots" and represent sites where minor accident levels cluster; gray points represent not significant points. Different shades of color correspond to different confidence levels, and the darker the color is, the higher the confidence level can be. Major accidents usually occur in the subareas including the Ningde sea area and the Fuzhou sea area, while minor accidents occur in the subareas including the Zhangzhou sea area, Xiamen sea area, and Quanzhou sea area. The gray points are distributed in the Putian sea area and part of the Fuzhou sea area, indicating that the characteristics of accidents in this sea area are not significant. Figs. 9 and 10 suggest that the results of the cold/hot spot analysis are consistent with those of the Local Moran's

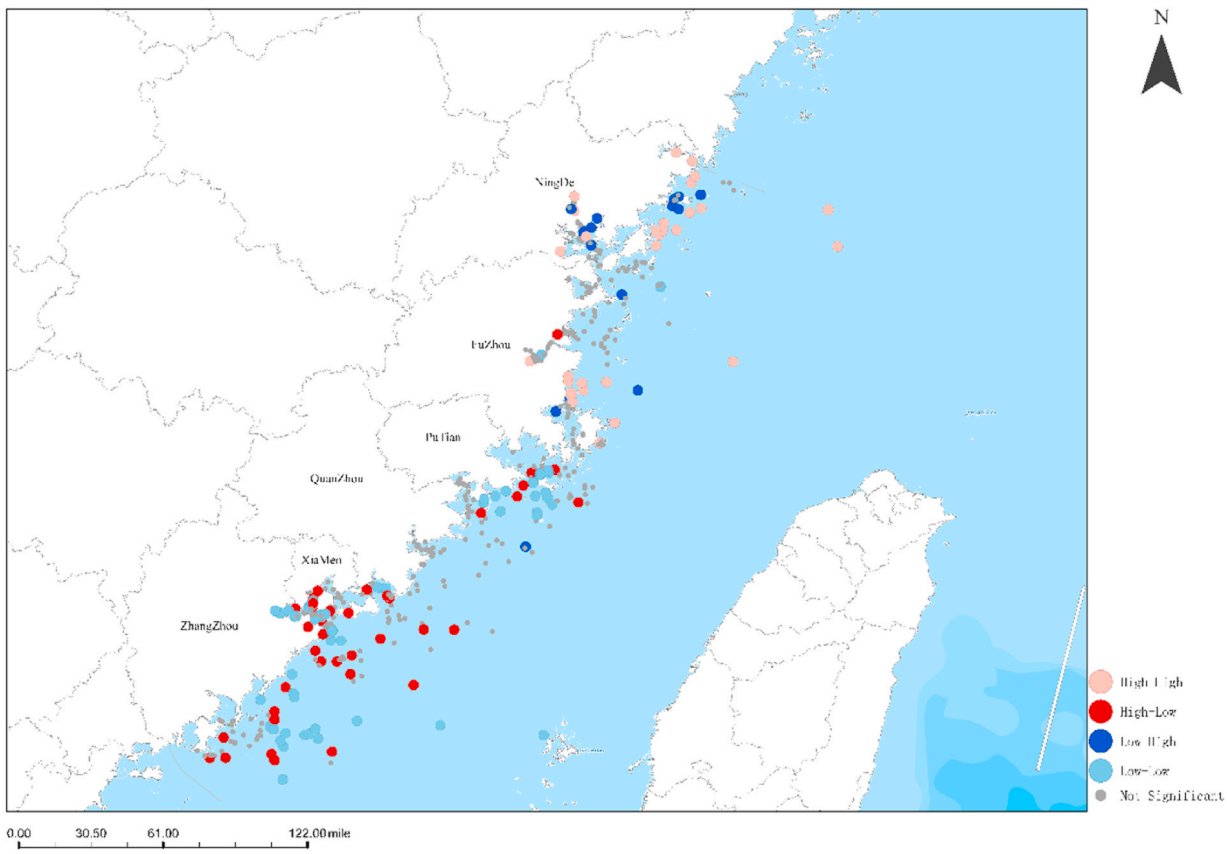

Fig. 9. Anselin local Moran's I.

I.

4.4. Accident prediction model

The risk of maritime accidents is closely related to traffic characteristics such as the speed, size, and regional flow of ships (Gan et al., 2014; Zhang et al., 2022). Since there are few accident spots at sea and they have obvious dispersed distribution characteristics, in this paper, the mapping relationship between the number of accidents and the traffic characteristics of small grid areas is established within a grid by unifying the grid scale using standardized grid statistics. In this paper, we sample characteristic dimensions for accurate results in Fig. 11: *a* represents the characteristics of the ship's mean length, *b* represents the standard deviation characteristics of the course, *c* represents the standard deviation characteristics of the ship width, *d* represents the characteristics of the ship's mean speed, *e* represents the standard deviation characteristics of the ship speed, *f* represents the standard deviation characteristics of the ship length, *g* represents the characteristics of the ship's mean width, and *h* represents the characteristics of the ship flow. These dimensions were chosen for the following reasons:

4.4.1. Static scale dimensions

① The ship's length and width reflect the static traffic characteristics of the ship in the grid. The larger the length and width, the more the grid is favored by large ships passing by. Small characteristic data indicate that these grids are not on the main route of large ship traffic, and the risk of accidents decreases accordingly.

② Standard deviations of ship length and width are the deviations of the static characteristics (length and width) of ships in the statistical sampling grid. The larger the characteristic value, the more significant the scale deviation of ships. Ships of different sizes may have their own unstandardized behaviors;

4.4.2. Traffic flow dimensions

① The average speed is the average speed of trajectory in the grids. This feature indicates that the grid area is within the main navigation channel and that large ships come and go frequently.

② Standard deviation is the deviation between the ship's speed and the mean value in the statistical sampling grid. The larger the standard deviation is, the greater the speed difference is, a trend observed in large and small ships;

③ Flow characteristics describe the size of the flow of ships passing through the sampling grid. Larger values indicate habitual travel by ships in this area and a possibility of accident risk.

4.4.3. Convergence risk dimensions

The standard deviation of the ship's course is the deviation between the course and the mean value in the statistical sampling grid. The larger the standard deviation is, the greater the difference between the course of ships in the grid area, signifying the convergence of courses.

(1) A standardized grid is established to divide the maritime research area into fine-grained grids.

(2) As maritime accidents are closely related to the size, speed, encounter situation, and flow of ships in the sea area, the grid statistical dimensions should include the ship flow through the grid, the mean length and width of the ship, the standard length and width of the ship, the standard deviation of the course and speed and the mean difference, as well as the standard deviation of heading and the mean difference.

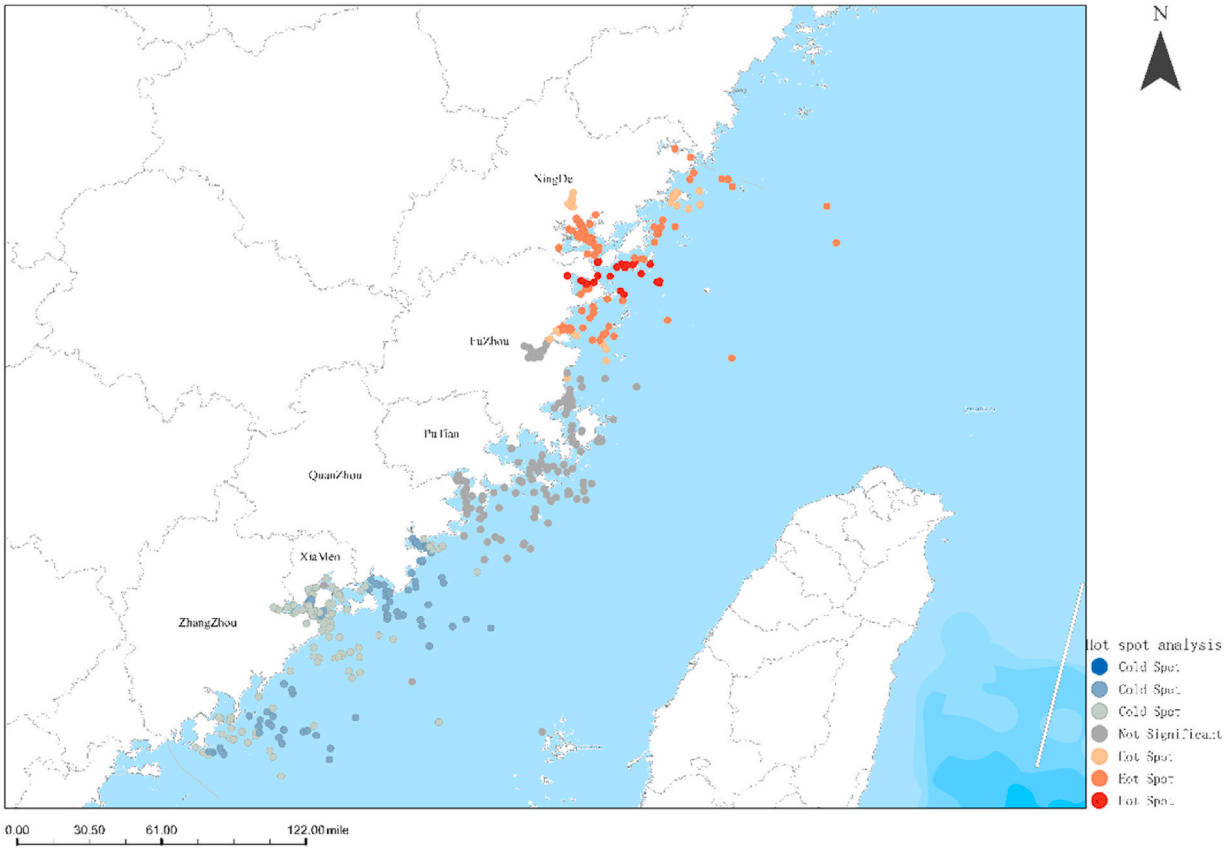

Fig. 10. Getis-Ord Gi* analysis.

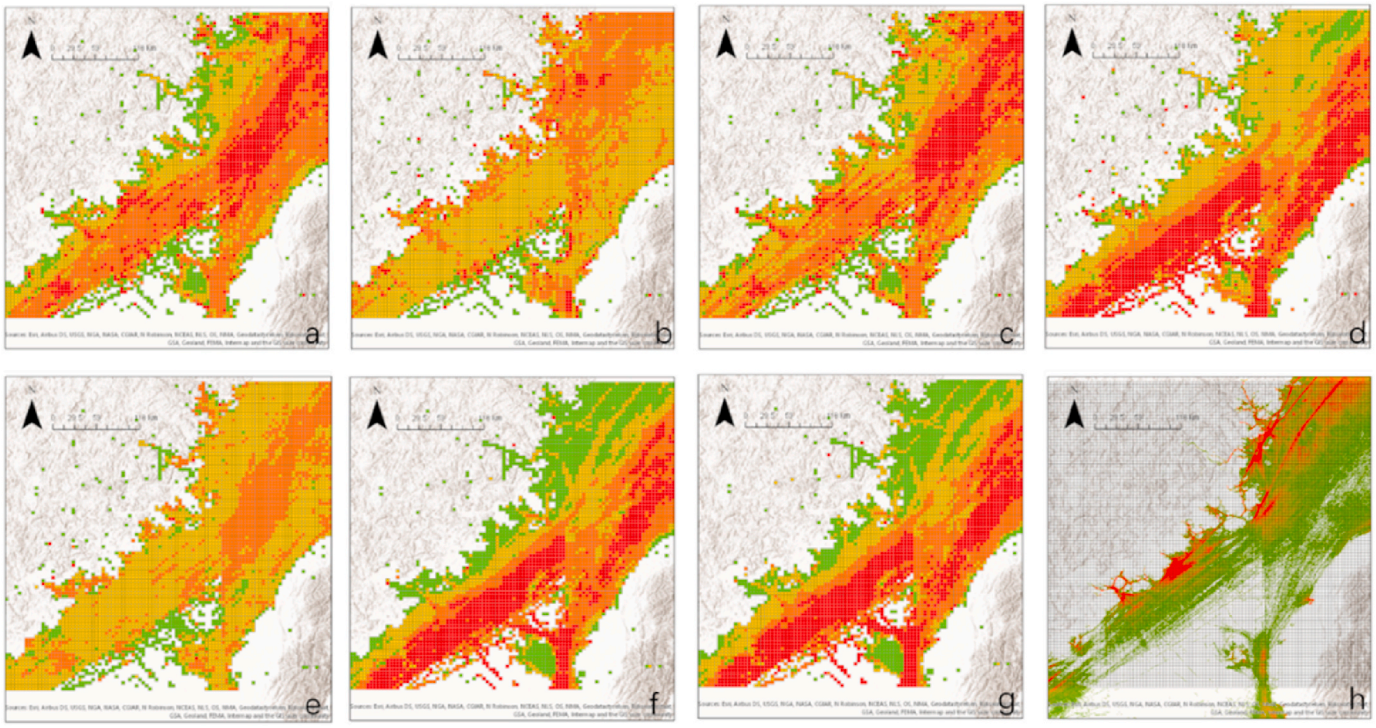

Fig. 11. Traffic flow characteristics and sea area gridding. The specific experimental steps in this paper are as follows.

(3) We carry out comparative experiments on the RF model, Adaboost model, GBDT model, Stacking combined model, CNN model, LSTM model, and SVM model, train each model with marine accident data, establish binary and multi-classification prediction tasks, complete the construction of the prediction model, and compare marine accident prediction performance of different integrated learning algorithms and classical prediction models.

4.4.4. Binary classification prediction model on "whether a grid area is an accident-prone area"

In this paper, AIS data are used to produce statistics on the traffic conditions in this area. A 100*100 standardized statistical grid is established in this sea area. The dimensions within the grid are the traffic flow in the grid, the mean length and width of the ship, the standard length and width of the ship, the standard deviation of the course, as well as the standard deviation of the speed and its mean value. During the process of training sample selection, the grid of "accidents occurred" is marked as the negative sample, and that of "no accidents occurred" is the positive sample. An extreme category imbalance between positive and negative samples exists, with the number of negative samples being only 295 and positive samples 9705 in 10,000 grids. Some positive samples are randomly selected from 9705 pieces of data to avoid overfitting of the model to the positive sample data. The ratio of positive and negative samples reaches 10:1, which facilitates the learning of negative sample features for the model. This part of the data is divided into training set and test set with a ratio of 8:2. Finally, the data sets are input into the abovementioned models and the prediction results of each model on "whether a grid area is an accident-prone area" are obtained.

The receiver operating characteristic curve (ROC) is taken as the evaluation index for four models. The ROC curve is commonly used to measure the accuracy of classification. The closer the ROC value is to 1, the better the model's classification effect is. The final results of the binary classification prediction model are shown in Fig. 12. The ROC integral values of the Random Forest model, Adaboost model, GBDT model, and Stacking combined model were 0.77, 0.75, 0.77, and 0.77. The result indicates that except for the Adaboost model, the results of other models have reached a high level of classification accuracy, satisfying the need for binary classification prediction of maritime accidents.

In order to verify the optimality of the combined model, this paper further compares the experimental results with those of classical prediction models (CNN, LSTM, and SVM). Among them, the CNN adopts the Resnet18 network model, characteristic of simplicity and practicality (Liu et al., 2021). The experimental comparison results are shown in Fig. 13. The ROC values of SVM model, LSTM model, and Resnet18 model are 0.76, 0.59, and 0.75 respectively. Thus, the integrated learning combined model is the optimal model.

For further comparisons, this paper takes accuracy, precision, recall, and F1 value as evaluation indexes to determine the optimization model. Accuracy is the proportion of positive and negative categories that are predicted accurately. Precision indicates the accuracy of the guess. Recall represents how many samples that are actually positive are predicted to be positive. F1 value is an evaluation index that can reflect both precision and recall. According to Table 4, the prediction effect of the Stacking combined model is the best, followed by the GBDT model and Random Forest model, and the last is the Adaboost model. Most classical prediction models are inferior to ML models.

Fig. 14 is the predicted grid diagram of "whether a grid area is an accident-prone area" in the Fujian sea area. The red grid represents "the area where the accident may occur", and the blank grid represents "areas where no accident may occur". The figure shows that risky areas are mostly located in coastal areas and estuaries. According to accident prediction results, the distribution of possible accident areas in the whole sea area is relatively uniform.

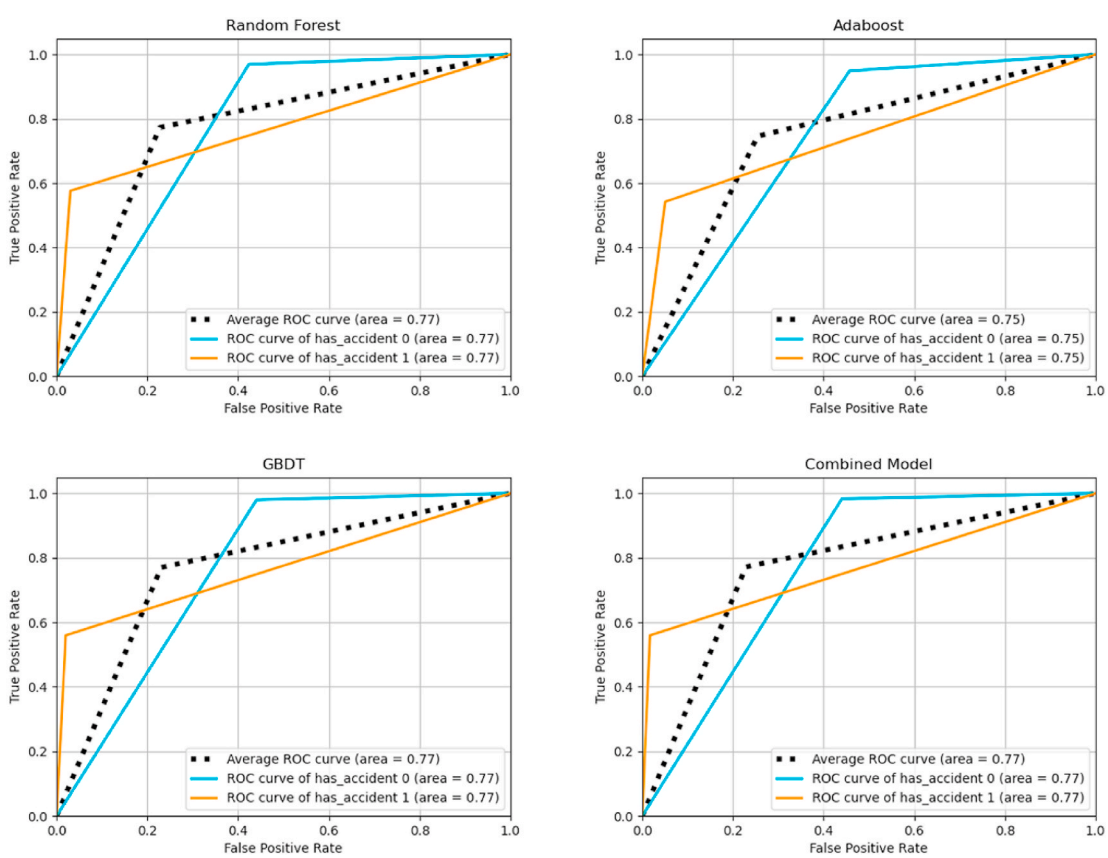

Fig. 12. Binary classification prediction model–ROC curve.

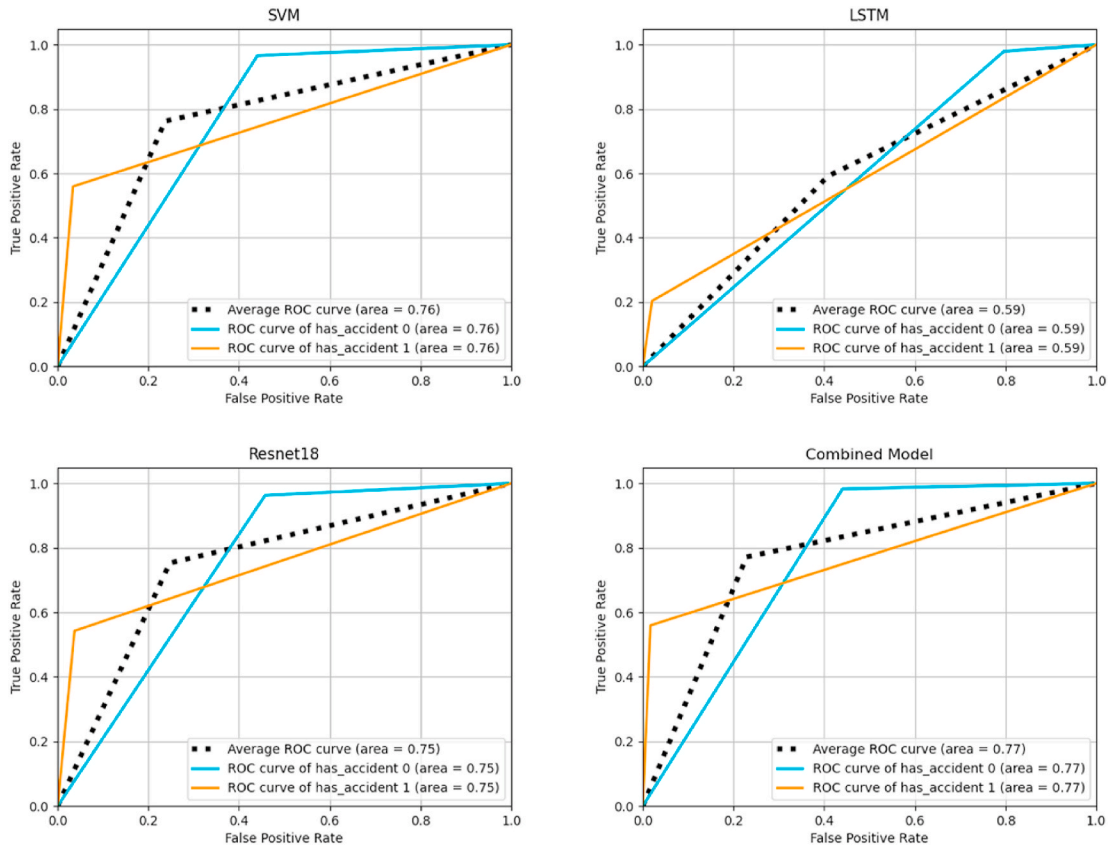

Fig. 13. Binary classification prediction model–ROC curve.

Table 4
Binary classification prediction model–Evaluation index of the prediction effect.

Evaluation Index	Random Forest	Adaboost	GBDT	SVM	LSTM	CNN (Resnet18)	Stacking combined model
Accuracy	0.904	0.881	0.901	0.898	0.850	0.893	**0.912**
Precision	0.898	0.874	0.906	0.892	0.828	0.885	**0.910**
Recall	0.904	0.881	0.910	0.898	0.850	0.893	**0.912**
F1	0.898	0.876	0.902	0.892	0.815	0.886	**0.904**

4.4.5. Multi-classification prediction model on "the accident severity"

This paper used AIS data to make statistics on the traffic conditions in this area. The sea area was gridded to 100*100 in the multi-classification preprocessing. The dimensions within the grid are the traffic flow in the grid, the mean length and width of the ship, the standard deviation of the course and speed and the average speed. When multiple labels are set, different levels correspond to different severity degrees of maritime accidents: 1.0 for minor accidents, 5.0 for ordinary accidents, 10.0 for major accidents, and 20.0 for serious accidents. In this paper, 295 pieces of accident data with AIS information are collected, among which the number of different types of accidents differs, with a ratio of 146:81:49:19. In order to avoid mistakes for the model learning caused by this imbalance, each type of data is interpolated linearly with the ratio of each category reaching 1:1:1:1 and 584 pieces of training data are obtained. Training set and test set are determined according to those data with a ratio of 8:2. Finally, the data sets are input into the above mentioned models, and the prediction results of each model are obtained.

Similar to the experimental steps of the two-classification prediction model, the final results of the multi-classification prediction model are shown in Table 5, Table 6, Figs. 15 and 16. The average ROC curve values of the Random Forest model, Adaboost model, GBDT model, Stacking combined model, SVM model, LSTM model, CNN model (Resnet18) were 0.74, 0.64, 0.67, 0.77, 0.71, 0.58, and 0.76. In addition, their corresponding F1 values were 0.725, 0.442, 0.693, 0.746, 0.568, 0.325, and 0.633, suggesting that the Stacking combined model has the best effect and can achieve a higher classification accuracy level, thus suitable for predicting the severity of maritime accidents.

Fig. 17 is the predicted grid diagram of "the accident severity" in the Fujian sea area. The green grid represents "minor accidents", the orange grid represents "ordinary accidents", the blue grid represents "major accidents", and the red grid represents "major accidents". Obviously, the green grid is widely distributed, which can be found in all sea areas of Fujian Province, followed by the orange grid. The blue grid is mainly distributed in subareas like the Ningde sea area, Fuzhou sea area, and Xiamen sea area, while the red grid is in the subareas like the Ningde sea area and the Fuzhou sea area.

The prediction results of "whether a grid area is an accident-prone area" and "the accident severity" in the Fujian sea area is obtained. In terms of traffic flow characteristics, the whole sea area has heavy traffic, dense routes, diverse types of ships with complex route conditions, and thus has a higher possibility of accidents than the sea area with sparse routes. In terms of the spatial distribution characteristics of accidents, the possible accident area is consistent with the distribution of accident kernel density and high-high and low-low clusters of accidents. For example, the area where a major accident is likely to occur coincides

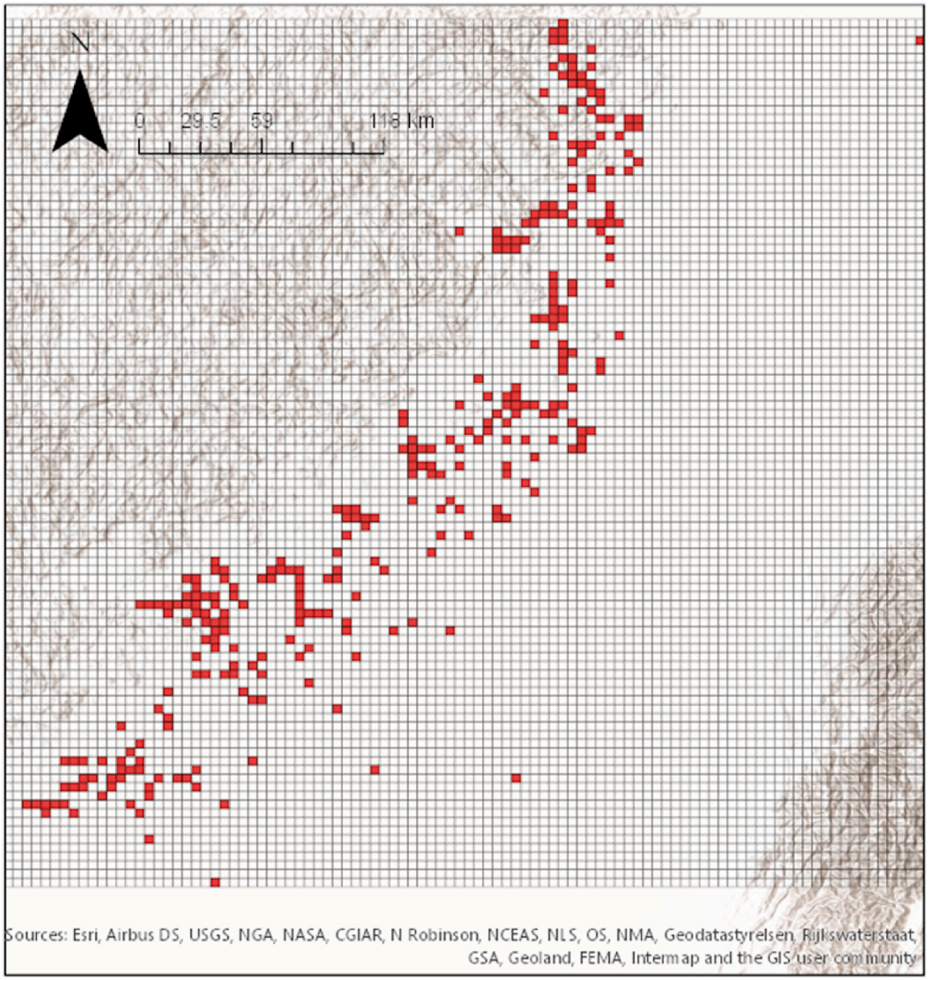

Fig. 14. The predicted grid diagram of "whether a grid area is an accident-prone area" in the Fujian sea area.

Table 5
Comparing the ROC curve value.

Type	Random Forest	Adaboost	GBDT	SVM	LSTM	CNN (Resnet18)	Stacking combined model
Minor accident	0.68	0.54	0.61	0.58	0.57	0.67	**0.66**
Ordinary accidents	0.81	0.57	0.77	0.68	0.62	0.68	**0.84**
Major accidents	0.81	0.65	0.78	0.83	0.62	0.78	**0.82**
Serious accidents	0.94	0.65	0.91	0.74	0.50	0.90	**0.94**
Average ROC curve	**0.81**	**0.68**	**0.77**	**0.71**	**0.58**	**0.76**	**0.82**

Table 6
Multi-classification prediction model–Evaluation index of the prediction effect.

Evaluation Index	Random Forest	Adaboost	GBDT	SVM	LSTM	CNN (Resnet18)	Stacking combined model
Accuracy	0.729	0.444	0.694	0.579	0.400	0.641	**0.750**
Precision	0.723	0.462	0.693	0.572	0.347	0.635	**0.745**
Recall	0.729	0.444	0.694	0.579	0.400	0.641	**0.750**
F1	0.725	0.442	0.693	0.568	0.325	0.633	**0.746**

with the area where there is a high risk of accident kernel density.

5. Conclusion and discussion

Maritime security in the Fujian Sea is of vital importance because it is an important international transport channel. This paper studies the geographic characteristics of maritime accidents and the relationship between characteristics of normal trajectories and accidents. Our paper proposes a GIS-based accident analysis framework to characterize the spatial distribution of traffic accidents based on accident data by employing kernel density analysis and spatial autocorrelation techniques. The correlation between accident occurrence and geographical spatial location, as well as the spatial clustering of accidents are analyzed. The studied sea area is gridded into several subareas based on maritime traffic characteristics using AIS data. Whether a grid area is an accident-prone area and the severity of accidents within the grid are

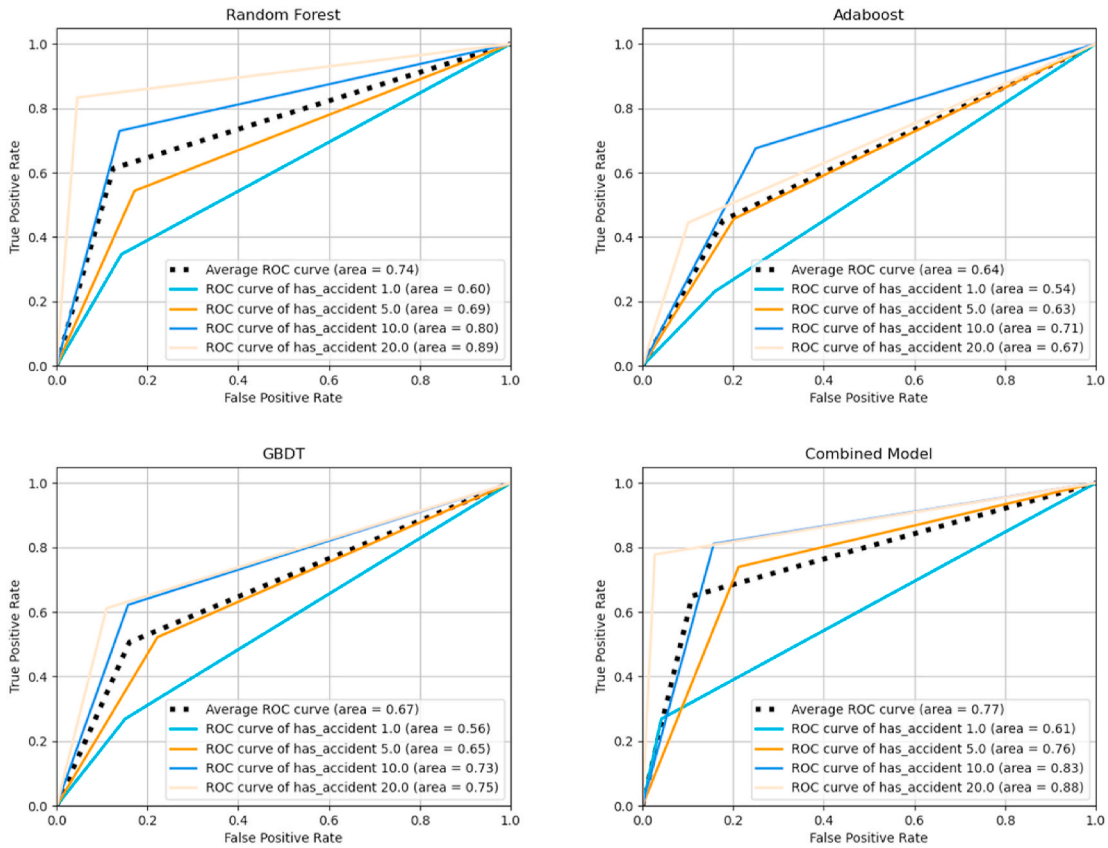

Fig. 15. Multi-classification prediction model–ROC curve.

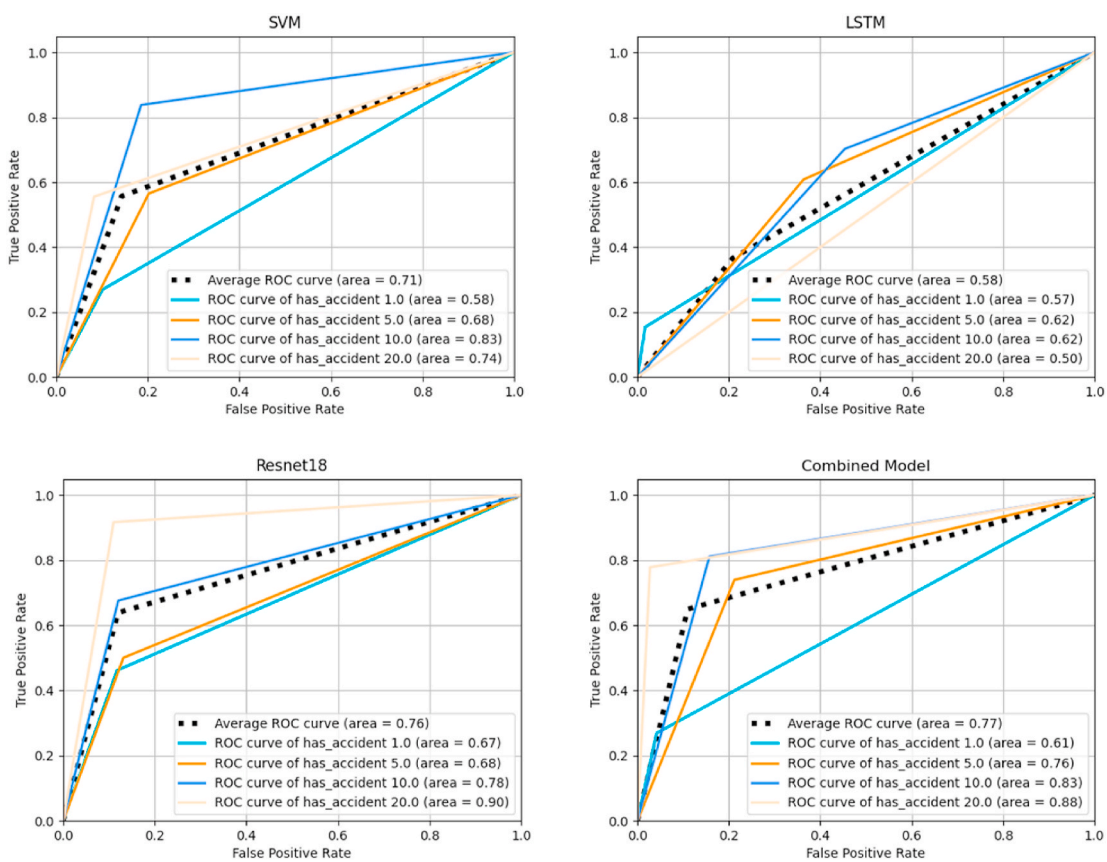

Fig. 16. Multi-classification prediction model–ROC curve.

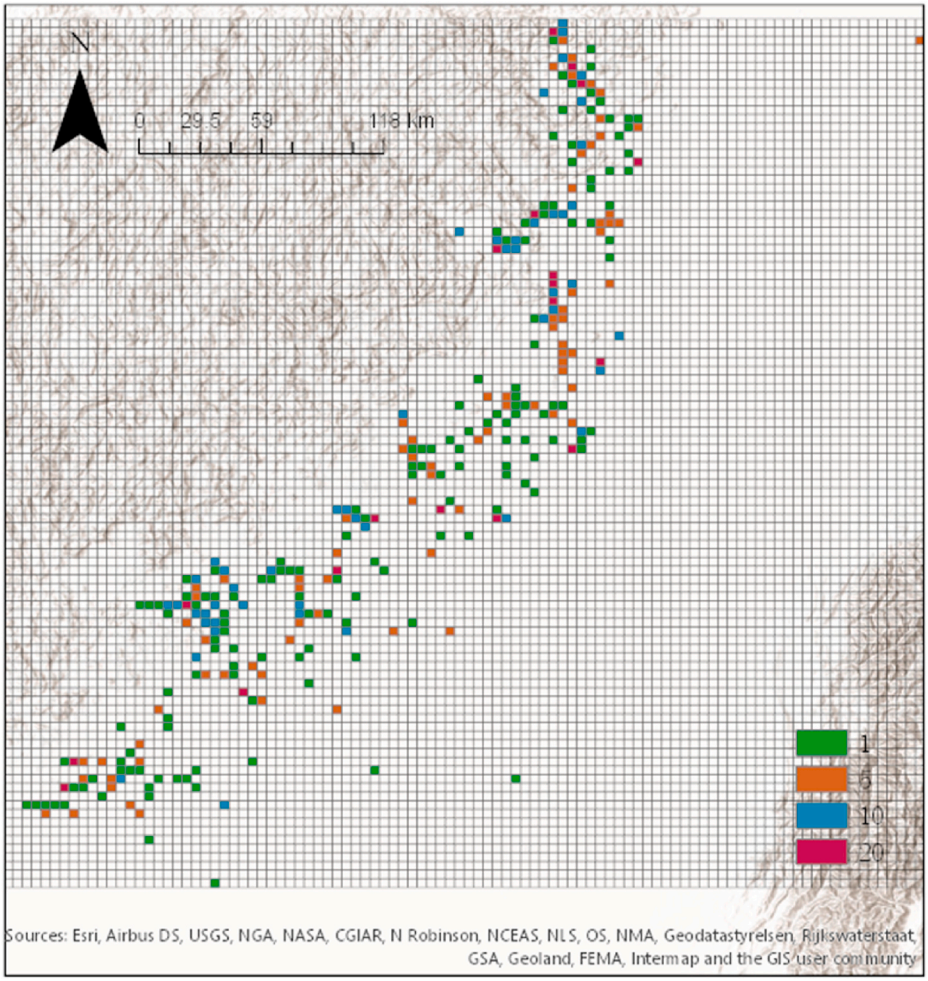

Fig. 17. The predicted grid diagram of "the accident severity" in the Fujian sea area.

evaluated and predicted by different ML models. The results of cluster analysis can provide richer spatial characteristics of maritime accidents. It can offer spatial patterns of accident severity and identify maritime accident-prone and high-risk areas with high accident severity, which cannot be achieved through traditional statistical analysis. A uniform grid scale for the Fujian sea area is established using standardized grid statistics. Analysis of the relationship between the number of accidents and the traffic characteristics of the grid areas using different ML models, including the Random Forest model, Adaboost model, GBDT model, and Stacking combined model, is compared with one traditional model (SVM) and deep learning models (CNN and LSTM). According to the analysis, the Fujian sea area shows typical cluster characteristics and a positive spatial correlation. In other words, the kernel density estimation indicates that subareas, including the Ningde, Fuzhou, and Xiamen subareas, generally have high accident density and the highest risk value within the whole Fujian sea area. High-high accident clustering occurs mainly in the Ningde and Fuzhou subareas, while the Xiamen, Putian, and Zhangzhou subareas have low-low clustering. In terms of prediction, the Stacking combined model outperforms the others with its high accuracy, precision, recall, and F1-score values of 0.912, 0.910, 0.912, and 0.904 in predicting whether a grid area is an accident-prone area and 0.750, 0.745, 0.750, and 0.746 in predicting the accident severity in the grid, indicating its superior maritime traffic accident prediction performance.

From the perspective of management, this paper uses real traffic trajectory data to establish an accident analysis and research framework based on GIS, and verifies the maritime space distribution characteristics and accident prediction conclusions of the Fujian sea area. This conclusion not only helps local maritime management personnel understand the spatial distribution of maritime accidents but can also be implemented in other sea areas for modeling, analysis, and prediction of maritime accidents. The new model proposed in this paper provides more detailed information to help competent authorities and stakeholders in the industry to supervise ship condition management and to formulate relevant policies to ensure maritime safety. The results can assist maritime administrators in centralizing regulatory forces to improve regulatory efficiency. For instance, through the kernel density analysis, the Ningde sea area, Fuzhou sea area, and Xiamen sea area are found to have high accident density and the highest risk value within the whole Fujian sea area. The Ningde port area has a high risk value for several reasons: the large area span in the port, the complex terrains, and the diverse natural conditions, especially wind, wave, current, and siltation. Ningde port is one of the busiest ports in Fujian Province, which has a huge daily traffic flow. Moreover, the ship tracks are interlaced and complex, and fishing boats are all over, but the channels are deep and narrow. The Fuzhou sea has a high risk value because the port area is mainly developed for passenger transport to Taiwan, in addition to freight, and dense routes and various types of ships complicate the transportation environment. The high risk value of the Xiamen sea area arises from geographical and environmental factors, hydrometeorological factors, and the high density of ship traffic. Through the spatial autocorrelation analysis method, the marine accident data set of the whole Fujian sea can be clustered with a spatial positive correlation pattern. High-high clustering occurs in the Ningde sea area and Fuzhou

sea area, and low-low clustering of accidents occurs in the Xiamen sea area, Putian sea area, and Zhangzhou sea area. Major accidents usually occur in the Ningde sea area and the Fuzhou sea area, while minor accidents are more likely to occur in the Zhangzhou sea area, Xiamen sea area, and Quanzhou sea area. Grid division can provide specific focuses for maritime supervision and reduce the workload of supervisors. The results indicate that some grids in the Ningde sea area, Fuzhou sea area, and Xiamen sea area have high accident density and high risks for accidents. Maritime administrators need to focus on strengthening the monitoring of cruise ships and unmanned aerial vehicles (UAVs) in these areas.

This study reveals potential causal relationships between traffic flow trajectories and maritime accidents. Trajectory characteristics were used to make predictions using machine learning. Thus, the prediction performance is based on the size of the collected data. Without effective data collection, it is difficult to achieve accurate predictions with this method. Therefore, data volume was a major limitation in this study. Substantial maritime trajectory data needs to be available for effective predictions. The spatiotemporal distribution characteristics of maritime traffic accidents should be investigated based on multi-source data, such as fusing occurring time of accidents, in future studies, and spatiotemporal accident prediction models should be established.

CRediT authorship contribution statement

Yang Yang: Conceptualization, Methodology, Writing – original draft. **Zheping Shao:** Writing – original draft. **Yu Hu:** Software. **Qiang Mei:** Conceptualization. **Jiacai Pan:** Data contribution. **Rongxin Song:** Writing – original draft. **Peng Wang:** Conceptualization.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

References

Ali, A., Ana, M.R.A., Santos, S.C.G.M., 2021. Comparison of Multivariate Regression Models and Artificial Neural Networks for Prediction Highway Traffic Accidents in Spain: A Case Study. Transport. Res. Procedia 58 (2352–1465), 277–284.

Anselin, L., 1995. Local indicators of spatial association—LISA. Geogr. Anal. 27 (2), 93–115.

Artyomov, E., Yadid-Pecht, O., 2005. Modified high-order neural network for invariant pattern recognition. Pattern Recogn. Lett. 26 (6), 843–851.

Bai, Y., Liu, K., Wang, Y.Y., 2022. Comparative analysis of thermal preference prediction performance in different conditions using ensemble learning models based on ASHRAE comfort database II. Build. Environ. 223, 109–462.

Blanco, V., Japón, A., Puerto, J., 2022. A mathematical programming approach to SVM-based classification with label noise. Comput. Ind. Eng. 172, 108–611.

Chai, T., Xue, H., Sun, K.B., Weng, J.X., 2020. Ship accident prediction based on improved quantum-behaved PSO-LSSVM. Math. Probl Eng. 2020 (8823322).

Chen, R.Q., Zhang, C.G., Xu, b., Zhu, Y.H., Zhao, F., Han, S.Y., Yang, G.J., Yang, H., 2022. Predicting individual apple tree yield using UAV multi-source remote sensing data and ensemble learning. Comput. Electron. Agric. 201, 107–275.

Crimmins, M., Park, S., Smith, V., Smith, V., Kremer, P., 2021. A Spatial Assessment of High-Resolution Drainage Characteristics and Roadway Safety during Wet Conditions, 133. Applied Geography, pp. 102–477.

Dulebenets, M.A., 2018. A comprehensive multi-objective optimization model for the vessel scheduling problem in liner shipping.Int. J. Prod. Econ. 196, 293–318.

Fang, X.Q., Guo, X.M., Yuan, L., 2021. Application of random forest algorithm in global drought assessment. J. Geo-inform. sci. 23 (1040–1049), 200–474.

Feizizadeh, B., Omarzadeh, D., Sharifi, A., Rahmani, A., Lakes, T., Blaschke, T., 2022. A GIS-based spatiotemporal modelling of urban traffic accidents in tabriz city during the COVID-19 pandemic. Sustainability 14 (12), 74-68.

Gan, L.X., Zhang, L., Zou, Z.J., Wen, Y.Q., Zhang, H., 2014. Analysis of Vessel Traffic Filw Based on Field Method. J. Shanghai Jiao Tong Univ. 551–557.

Getis, A., Ord, J.K., 2010. The analysis of spatial association by use of distance statistics. Geographical analysis.Perspect. Spat. Data Anal. 24 (3), 189–206.

Gu, J.X., Wang, Z.H., Kuen, J., Ma, L.Y., Shahroudy, A., Shuai, B., Liu, T., Wang, X.X., Wang, G., Cai, G.F., Chen, T.H., 2018. Recent Advances in Convolutional Neural Networks. Pattern Recogn. 77, 354–377.

Guo, Q.W., Guo, B.H., Wang, Y.G., Tian, S.X., Chen, Y., 2022. A combined prediction model composed of the GM (1,1) model and the BP neural network for major road traffic accidents in China. Math. Probl Eng. (8392759).

Hammami, M.A.I., Matisziw, T.C., 2021. Measuring the Spatiotemporal Evolution of Accident Hot Spots, 157. Accident Analysis & Prevention, pp. 106–133.

Kalantari, M., Shahraki, S.Z., Yaghmaei, B., Ghezelbash, S., Ladaga, G., Salvati, L., 2021. Unraveling urban form and collision risk: the spatial distribution of traffic accidents in Zanjan, Iran. Int. J. Environ. Res. Publ. Health 23, 44–98.

Katanalp, B.Y., Ezgi, E., 2021. GIS-based assessment of pedestrian-vehicle accidents in terms of safety with four different ML models. J. Transport. Saf. Secur. 14, 1–40.

Kim, J.H., Kim, J., Lee, G., Park, J., 2021. Machine learning-based models for accident prediction at a Korean container port. Sustainability 13 (16).

Kumar, M.B., Debasish, J., Niva, M., Somula, R., Rawal, B.S., Hasmat, M., Gopal, C., Smriti, S., 2022. Machine learning based accident prediction in secure IoT enable transportation system. J. Intell. Fuzzy Syst. 42, 713–725.

Li, Q.F., Song, Z.M., 2022. High-performance concrete strength prediction based on ensemble learning. Construct. Build. Mater. 324, 126–694.

Li, L.C., Sheng, X., Du, B., Wang, Y.G., Ran, B., 2020. A deep fusion model based on restricted Boltzmann machines for traffic accident duration prediction. Eng. Appl. Artif. Intell. 93, 103–686.

Li, X., Liu, Y., Fan, L.S., Shi, S.L., Zhang, T., Qi, M.J., 2021. Research on the prediction of dangerous goods accidents during highway transportation based on the ARMA model. J. Loss Prev. Process. Ind. 72, 104–583.

Li, Xy, Cheng, K., Tan, S.C., Huang, T., Yuan, D.D., 2022. Fault diagnosis method of nuclear power plant Based on Adaboost. Nucl. Power Eng. 2020, 1–9.

Liang, M.H., Liu, W., Li, S.C., Xiao, Z., Liu, X., Lu, F., 2021. An unsupervised learning method with convolutional auto-encoder for vessel trajectory similarity computation. Ocean Eng. 225, 108–803.

Lin, Y.L., Li, R.M., 2020. Real-time Traffic Accidents Post-impact Prediction: Based on Crowdsourcing Data. Accid. Anal. Prev. 145, 105–696.

Liu, Y., She, G.R., Chen, S.X., 2021. Magnetic resonance image diagnosis of femoral head necrosis based on ResNet18 network. Comput. Methods Progr. Biomed. 2021, 106–254.

Luo, J., Sun, H., Zhang, W.P., 2022. The Scheme of Re-floating a Grounded Vessel and Risk Analysis Based on M.V. EVER GIVEN. Am. J. Traffic Transport. Eng. 7, 51–55.

Ma, Q.L., Huang, G.H., Tang, X.Y., 2021a. GIS-based analysis of spatial-temporal correlations of urban traffic accidents. Eur. Transport Res. Rev.

Ma, Z.J., Mei, G., Salvatore, C.M., 2021b. An Analytic Framework Using Deep Learning for Prediction of Traffic Accident Injury Severity Based on Contributing Factors. Accident Analysis and Prevention.

Macwdo, M.R.O.B.C., Maia, M.L.A., Rabbani, E.R.K., Neto, O.C.C.L., Andrade, M., 2021. Traffic accident prediction model for rural highways in Pernambuco. Case Stud. Transport Pol. 10 (1), 278–286.

Misuk, L., Hyun, Y., Kwan, 2021. An Analysis on the Spatial Pattern of Local Safety Level Index Using Spatial Autocorrelation - Focused on Basic Local Governments, Korea, , 1st39. Jouranl of the Korean society of survey,geodesy,photogrammetry,and cartography, pp. 29–40.

MSA, 2010. Maritime Safety Administration of People's Republic of China. https://www.msa.gov.cn/page/article.do?type=hsfg&articleId.

Nermin, Hasanspahić, Vujičić, Srđan, Miho, Kristić, Mario, Mandušić, 2022. Improving safety management through analysis of Near-Miss reports—a tanker ship case study. Sustainability 14 (3), 031–094.

Ord, J.K., Getis, A., 1995. Local spatial autocorrelation statistics: distributional issues and an application. Geogr. Anal. 27, 286–306.

Pei, Z.H., Ge, M., Li, H., Wang, C.X., 2022. Environmental factors influencing HDL-C in middle-aged and elderly Chinese population based on random forest model. J. Geo-inform. sci. 7, 1286–1300.

Rong, H., Teixeira, A.P., Soares, C.G., 2021. Spatial Correlation Analysis of Near Ship Collision Hotspots with Local Maritime Traffic Characteristics. Reliab. Eng. Syst. Saf. 209, 107–463.

Shan, S., Li, C.X., Ding, Z.T., Wang, Y.Y., Zhang, K.J., Wei, H.K., 2022. Ensemble Learning Based Multi-Modal Intra-hour Irradiance Forecasting. Energy Convers. Manag. 270, 116–206.

Shen, J.Q., Wang, Q., Hu, S.S., Lu, J.S., Tian, Y.X., 2022. Prediction of Feature Size and Performance of Fe-36Ni/304L Lap Joint Based on GBDT Algorithm. J. Tianjin Univ. Sci. Technol. 55, 350–356.

Shu, Y., Daamen, W., Ligteringen, H., Hoogendoorn, S.P., 2017. Influence of external conditions and vessel encounters on vessel behavior in ports and waterways using Automatic Identification System data. Ocean Eng. 131, 1–14.

Shu, Y., Daamen, W., Ligteringen, H., Wang, M., Hoogendoorn, S.P., 2018. Calibration and validation for the vessel maneuvering prediction (VMP) model using AIS data of vessel encounters. Ocean Eng. 169, 529–538.

Silverman, B.W., B, E., 1998. Density Estimation for Statistics and Data Analysis, 1998. Routledge, New York, pp. 1–176.

Wang, H.X., Liu, Z.J., Liu, Z.C., Wang, X.J., Wang, J., 2022. GIS-based analysis on the spatial patterns of global maritime accidents. Ocean Eng. 245, 110–569.

Xing, X.Y., Yang, X.C., Xu, B., Jin, Y.X., Guo, J., Yang, D., Wang, P., Zhu, L.B., 2021. Remote sensing estimation of grassland aboveground biomass based on random forest. J. Geo-inform. sci. 23 (7), 1312–1324.

Xiong, H.J., Yi, H.J., Fu, L.Y., 2021. Traffic Safety Evaluation and Accident Prediction of Freeway: Evidence from China. Technical Gazette 28, 1904–1911.

Yan, W., 2020. Design of ship navigation trajectory analysis and application system based on image processing technology. J. Coast Res. 115, 211–213.

Yang, J.M., Yuna, N., Kwon, O.H., 2021. Analysis of the Characteristics of Marine Accidents Considering the Spatial Characteristics of Coastal Areas. Traffic Safety Research 40, 1–9.
Yang, Z.K., Zhang, W.P., Juan, F., 2022. Predicting multiple types of traffic accident severity with explanations: a multi-task deep learning framework. Saf. Sci. 146, 0925–7535.
Ye, Y., Liu, X.J., Zhang, Z., Li, Z.L., Hu, Y.Q., 2022. patiotemporal variations of chemical weathering intensity in large drainage basin and its potential climatic implications: A case study from the Yangtze River Valley. J. Geochem. Explor. 243, 107–093.
Yuan, C.F., Hu, Y.C., Zhang, Y.L., Zuo, T., Wang, J.H., Fan, S.J., 2021. Evaluation on consequences prediction of fire accident in emergency processes for oil-gas storage and transportation by scenario deduction. J. Loss Prev. Process. Ind. 72, 104–570.
Zhang, B.B., Wu, S., Cheng, S.F., 2019. Spatial Characteristics and Factor Analysis of Pollution Emission from Heavy-Duty Diesel Trucks in the Beijing–Tianjin–Hebei Region, China. Int. J. Environ. Res. Publ. Health 16, 817–819.
Yuan, Z., Liu, J.X., Liu, Y., Zhang, Q., Liu, W., 2020. A multi-task analysis and modelling paradigm using LSTM for multi-source monitoring data of inland vessels. Ocean Eng. 213, 107–604.
Zhang, Y., Sun, X.K., Chen, J.H., Cheng, C., 2021. Spatial Patterns and Characteristics of Global Maritime Accidents. Reliab. Eng. Syst. Saf. 206, 107–310.
Zhang, M.Y., Zhang, D., Fu, S.S., Kujala, P., Hirdaris, S., 2022. A Predictive Analytics Method for Maritime Traffic Flow Complexity Estimation in Inland Waterways. Reliab. Eng. Syst. Saf. 220, 108–317.