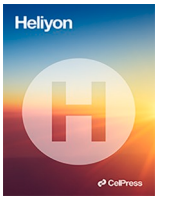
Heliyon 10 (2024) e30046

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Predicting types of human-related maritime accidents with explanations using selective ensemble learning and SHAP method

He Lan, Shutian Wang *, Wenfeng Zhang

School of Economics and Management, Dalian Ocean University, Dalian, 116023, China

ARTICLE INFO

Keywords:
Maritime accidents
Seafarers' unsafe acts
Risk prediction
Selective ensemble learning

ABSTRACT

Maritime accidents frequently lead to severe property damage and casualties, and an accurate and reliable risk prediction model is necessary to help maritime stakeholders assess the current risk situation. Therefore, the present study proposes a hybrid methodology to develop an explainable prediction model for maritime accident types. Based on the advantages of selective ensemble learning method, this study pioneers to introduce a two-stage model selection method, aiming to enhance the predictive accuracy and stability of the model. Then, SHAP (Shapley Additive Explanations) method is integrated to identify effective mapping associations of seafarers' unsafe acts and their risk factors with the prediction results. The results demonstrate that the model developed achieves good prediction performance with an accuracy of 87.50 % and an F1-score of 84.98 %, which benefits stakeholders in assessing the type of maritime accident in advance, so as to make proactive intervention measures.

1. Introduction

Maritime accidents are the unexpected and abnormal events of ships, often resulting in casualties and property losses [1]. Maritime safety has been a concern for maritime authorities since the beginning of shipping [2]. Depending on the characteristic of maritime accident, various types of maritime accident are defined. According to China Maritime Safety Administration (MSA) [3], there are six main maritime accident types, which are collision, grounding, sinking, contact, fire/explosion, and wind. To prevent maritime accidents, numerous endeavors have been made to devise countermeasures for improving maritime safety [4]. For instance, International Maritime Organization (IMO) proposed International Convention on Standards of Training, Certification and Watchkeeping for Seafarers (STCW Convention) and the International Safety Management Code (ISM Code), which promoted maritime safety from the perspective of seafarer training and management. However, maritime accidents are the results of coupling complex and uncertain risk factors [5], for this reason, the growth and breakthrough in maritime safety research have been relatively slow compared to the frequency of accidents [6,7]. Therefore, even minor improvements in maritime accident prevention would have a great positive impact on safety operation.

Numerous investigations have been devoted to analyzing the causes of accidents in the industry, ultimately pointing to unsafe acts as the leading causes of accidents [8,9]. The importance of controlling seafarers' unsafe acts was recognized by the international shipping industry as early as the 1970s [5], due to the general belief that maritime accidents are the direct result of seafarers' unsafe acts, which has been confirmed by several studies [10]. For instance, Wróbel [11] claimed that unsafe acts were responsible for 80 % of

* Corresponding author.
E-mail address: 202301073@dlou.edu.cn (S. Wang).

https://doi.org/10.1016/j.heliyon.2024.e30046
Received 5 December 2023; Received in revised form 17 April 2024; Accepted 18 April 2024
Available online 26 April 2024

maritime accidents, facilitating increased studies in this field [12,13]. After examining 540 maritime accident reports, Lan et al. [14] pointed out that different types of maritime accident include specific seafarers' unsafe acts. The main differences of unsafe acts between collisions and groundings are resource management deficiencies, communication failures, decision errors, skill-based errors, and violations [15]. Chauvin et al. [16] stated that decision errors were the typical causes of ship collisions, which was also confirmed by Graziano et al. [17]. Meanwhile, Yıldırım et al. [15] identified that bridge resource management deficiencies were the most frequent errors and prerequisites for the groundings. Therefore, developing prediction model of maritime accident types using seafarers' unsafe acts may be an effective way to prevent maritime accidents in advance.

Machine learning (ML) provides an effective method for solving multivariate, nonlinear and complex problems, and is widely utilized in several fields for risk prediction [18,19]. It has been suggested that ML would perform better than traditional statistical models [20]. In the transportation field, ML has been employed to predict the likelihood [21] and the severity of accidents [22,23]. However, in the maritime domain, ML has received little attention for risk prediction [24], while most studies focus on the causation analysis of ship accidents. For instance, Qiao et al. [5] employed artificial neural network (ANN) to assess human factors contribute to maritime accidents, and suggested that unsafe preconditions and unsafe supervision are the two primary considerations for human factors analysis, especially supervision failures by shipping companies and ship owners. Bye and Aalberg [25] applied the multinomial logistic regression model to identify risk indicators associated with groundings and collisions based on the Automatic Identification System (AIS) data and ship accident data in Norwegian waters. The results indicated that ship type and poor visibility would increase the likelihood of ship accidents. In addition, a novel research area is the development of anomaly detection models using ship traffic data and machine learning for real-time monitoring of ship risks [26]. Rawson et al. [27] developed extreme gradient boosting (XGBoost) model, random forest (RF) model, support vector machines (SVM) model, and logistic regression (LR) model to monitor the risk of maritime navigation under adverse weather conditions. Although the potential of ML in risk prediction has been recognized, relevant research in the field of maritime safety is still limited.

Ensemble learning is a significant research direction in the field of machine learning, which makes prediction results more reliable and accurate by combining several simple learners [28,29]. It could greatly enhance the generalization ability of the model and decrease computational errors created by a single ML model. Bagging and Boosting are two classical homogeneous ensemble learning algorithms, for example, random forest is a Bagging-based ensemble learning algorithm that aggregates the outputs from numerous decision trees to generate a final prediction result [30]. In contrast, Stacking ensemble learning algorithm uses heterogeneous learners to develop several independent models in parallel, on top of which the meta-learner is developed to achieve the aggregation of final results [31]. Compared with homogeneous ensemble learning algorithms, Stacking algorithm improves the diversity of models and could further enhance the generalization of models. Although ensemble learning models are advantages in conducting risk prediction, the prediction speed of ensemble models decreases significantly as the number of individual learner increases. Additionally, ensemble models may not always achieve satisfactory performance due to the involvement of several individual learners with poor performance [32]. As a result, Zhou et al. [33] proposed the concept of selective ensemble learning, suggesting that removing the individual learner with poor performance and selecting only some of them to build the ensemble models could obtain better prediction performance and improve the model generalization ability. Selective ensemble learning, a new ML technique with outstanding performance and promising future, has not been explored in the domain of maritime safety.

On the other hand, the explanation of prediction models is critical for stakeholders to extract relevant risk factors and implement appropriate management responses. Explain ability is not yet clearly defined from a mathematical perspective, but exists in the form of a theoretical concept. Miller [34] defined explain ability from a non-mathematical perspective as the extent to which people can understand the reasons of model decisions. Currently, the commonly used explanation approach is based on the model's own characteristics for model explanation. For example, Zhu et al. [19] used eight machine learning algorithms to construct a severity prediction model for construction accidents respectively, and then utilized Random Forest algorithm to assess the impact of different factors on accident severity, and the results showed that emergency management and safety training played an important role in the severity of construction accidents. Similarly, Xu and Luo [35] used Random Forest algorithm to construct an early warning model for air traffic controllers' unsafe acts, and pointed out that operational capability, technical environment, insufficient supervision, and mental state were more important in predicting errors, whereas insufficient supervision, organizational climate, and organizational processes were more important in predicting violations. However, model-based explanation methods cannot reveal the detailed interactions between the influencing factors. As a result, post-hoc machine learning model interpretation methods such as Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) have gained many attentions. Among them, LIME could only be used to analyze the influence of single factor on the prediction results, and SHAP could explain the potential correlation between factors. In addition, SHAP is able to visualize the detailed relationship between the prediction results and the influencing factors [36]. Kim and Kim [37] predicted the hazard level of extreme hot weather based on Random Forest algorithm, and by introducing SHAP method, demographic, socio-economic, and climatic sectors were identified as the most contributing factors to the prediction process. Yang et al. [38] used Extreme Gradient Boosting (XGBoost) and SHAP method to explore the relationship between built environment factors and the spatial distribution of truck crashes. The results showed that demographics, land use and road network factors were highly correlated with the spatial distribution of truck crashes.

With the aim of developing an explainable prediction model for maritime accident types, a two-stage selective ensemble learning integrated with SHAP method is proposed. Specifically, the present study develops multiple heterogeneous individual learners and optimizes the hyperparameter using random search and grid search. Then, on the basis of Stacking ensemble learning, a two-stage selective ensemble learning method is proposed to build prediction model of maritime accident types from the perspective of model accuracy and diversity. Through the model performance evaluation and robustness test, the accuracy, stability and generalization ability of the model developed are examined. Moreover, SHAP method is integrated to identify the effective mapping

association and the association strength between the model prediction results and seafarers' unsafe acts, which provide useful references for maritime risk assessment and tailored prevention of maritime accidents, so as to enhance the reliability of maritime operations.

The remainder of the present study is structured as follows. In Section 2, the proposed methodology is described. In Section 3, data source and the application of the proposed methodology are presented. Finally, Section 4 provides the discussion and Section 5 gives the conclusions.

2. Methodology

Fig. 1 illustrates the flowchart of the proposed methodology, in which selective ensemble learning is introduced to select individual learners based on the two-stage principle of diversity and accuracy, and Stacking-based ensemble learning method is adopted to develop the prediction model of maritime accident types.

2.1. Stacking ensemble learning method

Stacking utilizes the idea of hierarchical fusion to aggregate individual learners by meta-learner to improve model performance. To avoid over fitting problems, meta-learner is usually chosen as a simple-structured single machine learning model, therefore, this study selects SVM as the meta-learner. Stacking ensemble learning method consists of 2 main layers, and layer 1 consists of N heterogeneous base learners. First, the original dataset is separated into training set and test set, and the training set is K-fold divided. Then, the base learners in the layer 1 are trained for K times, and train the meta-learner by using the output of layer 1 as the input of layer 2. Finally,

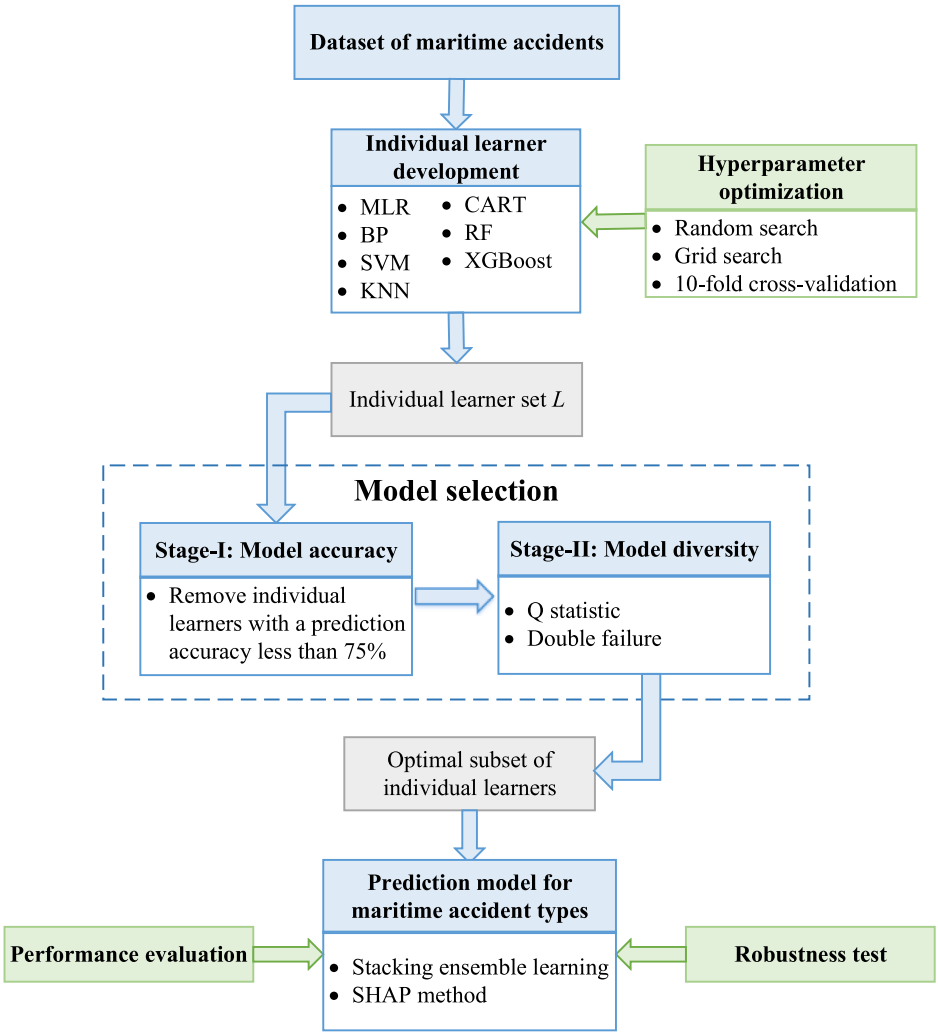


Fig. 1. The flowchart of the proposed methodology.

the meta-learner outputs the final prediction results. Stacking ensemble learning integrates the outputs of multiple heterogeneous learners to improve overall prediction performance and model generalization. Fig. 2 illustrated the process of stacking ensemble learning method.

2.2. Individual learners

In the present study, seven machine learning algorithms are determined to develop heterogeneous individual learners, including multinomial logistic regression (MLR), support vector machine (SVM), back propagation neural network (BP), K-nearest neighbour (KNN), classification and regression tree (CART), random forest (RF), and extreme gradient boosting (XGBoost). These algorithms are chosen because they have been applied in other transportation domains and have demonstrated good predictability. In addition, selective ensemble learning using individual learners with different structures could enhance generalization ability of the model developed. These seven machine learning algorithms are briefly described below.

Logistic regression is a statistical method for modeling the probability of a binary dependent variable. It assumes a linear relationship between the log odds of the dependent and independent variables. It has been applied in maritime risk assessment [39]. When response variables have more than two levels, MLR as the extension of logistic regression is utilized, and assumes that the categories of the dependent variables are completely independent. MLR predicts a different logistic regression model for every dummy variable, and every model has an individual collection of regression coefficients and intercepts. , which can be compared with the reference category to obtain results that predict the likelihood of success of the variable in that category [40].

Back Propagation (BP) neural network constructs complex relationships by introducing nonlinear transformations, belong to the commonly used neural network structures trained by back propagation [30]. Without disclosing the mathematical functions in advance, a BP network is capable of learning and storing a vast array of input-output pattern mapping interactions. Its learning strategy aims to minimize the sum of squared errors of the network by back propagating continuously adjusted weights and deviations using the fastest descent method. The output layer, hidden layer, and input layer make up BP neural network structure. This study constructs the single hidden-layer BP neural network model, which is split into two stages. The first stage is the forward propagation of the signal from the input layer through the hidden layer and finally to the output layer. The second stage is the backward propagation of the error from the output layer to the hidden layer and finally to the input layer, adjusting the weights and deviations from the hidden layer to the output layer and from the input layer to the hidden layer. The formula for forward propagation is expressed as Equation (1).

$$f(x) = F\left(\sum_{i=1}^{n}(w_i x_i) + b\right) \tag{1}$$

where w and b are the weights and deviations of neurons.

Meanwhile, SVM is a popular algorithm that uses kernel methods to efficiently process nonlinear data. It has the ability to handle nonlinear data and good classification accuracy [19]. The algorithm uses the optimal hyperplane to divide the data into several classes.

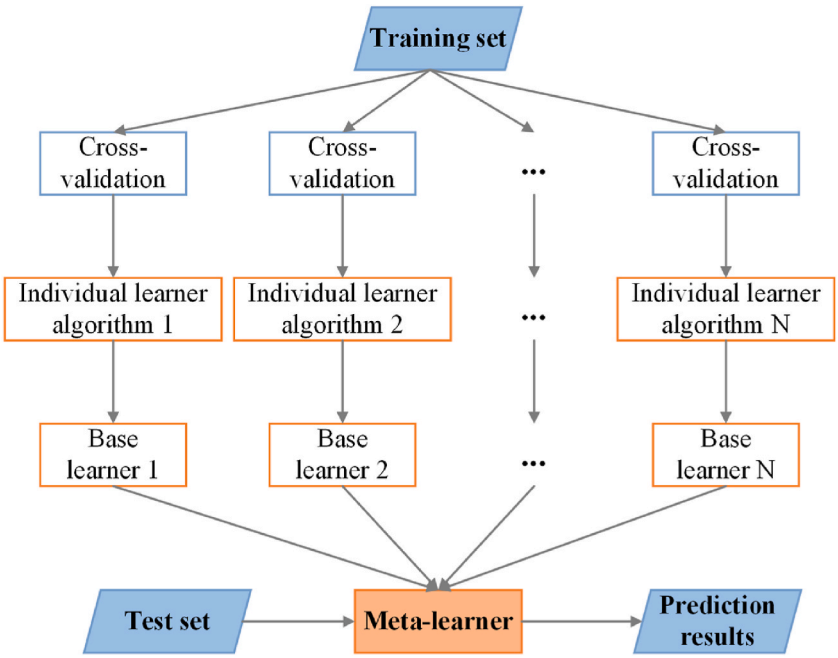


Fig. 2. The process of stacking ensemble learning method.

The hyperplane is essentially chosen to increase classification accuracy by taking into account the maximum margin of the closest point. If the training instances are denoted as (x_i, y_i) where $i = 1,2,\ldots N$, N represents the number of instances, y_i indicates the category of instances x_i in the training set. The Lagrange multiplier is used to compute the boundary function for the maximum margin by the pairwise formula as Equation (2).

$$Min\ L = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j a_i a_j k(x_i, x_j) \quad s.t. \quad a_i \geq 0; \forall i \text{ and } \sum_{i=1}^{N} a_i = 0 \tag{2}$$

KNN is one of the most straightforward classification algorithms. It achieves high classification accuracy in less computation time and solves the scalability problem [41]. All adjacent data points of the test tuple are identified and the distance between the training and test tuple is calculated using the Euclidean distance, which is defined by Equation (3), where X_i and Y_i (i = 1, 2, …,N) are the attributes of two samples/instances X and Y. The tuple with the minimum distance is then identified and the majority class label of the K nearest training tuples is assigned to it as the prediction class.

$$d(X, Y) = \sqrt{\sum_{i=1}^{N} (X_i - Y_i)^2} \tag{3}$$

CART is a powerful decision tree algorithm with the capacity to manage both categorical and continuous variables. The purpose of the algorithm is to split the dataset into two parts based on impurity measures, such as the Gini index. The Gini index of nodes can be expressed as Equation (4).

$$Gini(t) = 1 - \sum_{j=1}^{p} \left(\frac{n(j|t)}{n(t)}\right)^2 \tag{4}$$

where p is the number of classes of response attributes. The number of records in node t that belong to class j is indicated by A, while the total number of records in node t is indicated by B. The maximum impurity value of Gini index is 0.5 when each class is equally distributed. If the data in each class is equally distributed, the maximum impurity value of Gini index is 0.5. If there is a single class, the minimum impurity value of the Gini index is zero. The weighted average of Gini index for the descending nodes is computed as in Equation (5), to identify the attributes of the split

$$Gini(t)_{split} = \frac{n(t_L)}{n(t)} Gini(t_L) + \frac{n(t_R)}{n(t)} Gini(t_R) \tag{5}$$

where t_L and t_R denote the left and right child nodes of node t. The minimized attribute $Gini(t)_{split}$ is regarded as the root node and is selected for splitting. The recursive tree growth mechanism divides this root node into two branches. To avoid overfitting in this growth phase, CART employs pruning based on a minimal cost complexity criteria. The cost $(C_a(\mathrm{T}))$ of this pruning is allocated to each subtree, i.e., $C_a(T) = R(T) + aL(T)$, where $R(T)$ and $L(T)$ denote the ratio of training data misclassified by the tree T and the number of leaves T in the tree, respectively, and where $\mathrm{a}(\mathrm{a} \geq 0)$ is the complexity parameter. Reducing $C_a(T)$ is the aim of cost-complexity based pruning. When every leaf node in every branch of the tree has a projected class identified, the tree growth process eventually comes to an end.

RF is a tree-based ensemble classification algorithm that consists of a set of decision trees and uses bagging method to aggregate the decision trees [35]. It is simple to use and can handle over fitting problems. There are two primary stages in the construction of the RF model: the forest generating phase and the decision phase. Firstly, the random forest creates n CART by randomly dividing the training samples into n samples. After that, a majority vote is used to decide the ultimate classification outcome based on the classification of each individual decision tree.

With the use of different regularization strategies and management of the tree's complexity, XGBoost can generate results with increased accuracy [42]. Regression and classification issues can be resolved with the help of the built boosting trees through to its sophisticated methodology, which incorporates parallel tree boosting. Furthermore, the main element contributing to XGBoost's superiority is the learning process's goal function. A regularization term and a loss function make up the objective function. The regularization term limits the model's complexity and prevents overfitting; therefore the loss function computes the differences between each estimate and the actual value. Additionally, in order to minimize the objective function, XGBoost applies a second-order Taylor expansion on the loss function. As a result, the robust structure not only allows for quick computational processing but also more reliable outputs.

Each algorithm contains multiple hyperparameters that could be utilized to optimize model accuracy. Random search and grid search are two typical hyperparameter optimization methods, and both of them have strengths and weaknesses. Random search method selects hyperparameters randomly from the search space and tests the accuracy of the corresponding combinations by a performance estimation strategy, while grid search method is an exhaustive search method, in which the possible values of each hyperparameter are arranged and combined from the search space and evaluated by a cross-validation method to obtain the optimal learning algorithm. In the present study, the strengths and weaknesses of these two methods are complemented by first narrowing the hyperparameter search using random search and then adjusting the hyperparameters using grid search, combined with 10-fold cross-validation and 100 iterations to identify the set of hyperparameters with the best prediction accuracy.

2.3. Selective ensemble learning

Selective ensemble learning is the process of providing the learner selection phase between individual learners' development and aggregation. Through theoretical and experimental validation, selective ensemble learning has the potential to enhance the model's generalization performance, expedite prediction times, reduce overfitting problem, and minimize storage requirement [32]. After determining the individual machine learning algorithm and optimizing the hyperparameters on the training set to develop multiple individual learners with good prediction ability, the set of individual learners L is formed. Then, the selection of base learners for Stacking ensemble model is carried out to obtain the set of base learners E. In the present study, a two-stage selective ensemble method is proposed by considering the accuracy and diversity of models, and the steps are as follows:

S1: Remove the individual learners whose prediction accuracy does not exceed 75 %. The purpose is to avoid the negative impact of underperforming individual learners.

S2: Select the initial base learner E_1 from the set of individual learners L, and move it to the set of base learners E. The initial base learner is required to select the individual learner with the best prediction performance, taking into account the four-evaluation metrics: Accuracy, Precision, Recall, and F1-score.

S3: Select the second base learner E_2 from the set of individual learners L, and move it to the set of base learners E. The second base learner is required to select individual learner that diverges the most from E_1, taking into account the two-diversity metrics: Q statistics and the double failure (DF).

S4: Select the third base learner E_3 from the set of individual learners L, and move it to the set of base learners E. The third base learner is required to select the individual learner that diverges the most from the Bagging-based ensemble model constructed by E_1 and E_2.

S5: The process is repeated until L is reorganized into an alternative sequence in E.

S6: The top pre-defined individual learners that can optimize the prediction performance of the Stacking ensemble model are selected as the base learners.

2.3.1. Performance evaluation metrics

Confusion matrices are commonly used to evaluate the classification results of a test set by summarizing the learner results. Based on the confusion matrix, four typical evaluation metrics: accuracy, precision, recall, and F1-score can be computed to evaluate the predictive performance of the model. Table 1 shows an example of the confusion matrix to help understand the basic meaning of these metrics. Each column and row of the matrix represents a prediction class and an observation class, respectively. For example, N_{11} indicates the number of data that actually belongs to class 1 and the predicted outcome is also class 1, and N_{12} indicates the number of data that actually belongs to class 1 and the predicted outcome is class 2.

Accuracy is defined as the ratio of correctly predicted observations to the total observations (TN), which can be expressed as Equation (6).

$$Accuracy = \frac{\sum_{i=1}^{3} N_{ii}}{TN} \tag{6}$$

Precision is defined as the ratio of correctly predicted observations in the given class to all prediction observations in the same class, the precision value of class 1 is calculated by Equation (7), and the overall precision can be expressed as Equation (8)

$$Precision_1 = \frac{N_{11}}{\sum_{i=1}^{3} N_{i1}} \text{ (for class 1)} \tag{7}$$

$$Precision = \left(\frac{N_{11}}{\sum_{i=1}^{3} N_{i1}} + \frac{N_{22}}{\sum_{i=1}^{3} N_{i2}} + \frac{N_{33}}{\sum_{i=1}^{3} N_{i3}} \right) / 3 \tag{8}$$

Recall is defined as the ratio of correctly predicted observations under a given class to all reference observations in that class, the recall value of class 1 is calculated by Equation (9), and the overall recall can be expressed as Equation (10).

$$Recall_1 = \frac{N_{11}}{\sum_{j=1}^{3} N_{1j}} \text{ (for class 1)} \tag{9}$$

Table 1
Confusion matrix of three classes.

		Prediction		
		Class 1	Class 2	Class 3
Reference	Class 1	N_{11}	N_{12}	N_{13}
	Class 2	N_{21}	N_{22}	N_{21}
	Class 3	N_{31}	N_{32}	N_{33}

$$Recall = \left(\frac{N_{11}}{\sum_{j=1}^{3} N_{1j}} + \frac{N_{22}}{\sum_{j=1}^{3} N_{2j}} + \frac{N_{33}}{\sum_{j=1}^{3} N_{3j}} \right) / 3 \tag{10}$$

F1-score is an indicator calculated from recall and precision, which is expressed as Equation (11).

$$F1 - score = \frac{2 \times Precison \times Recall}{Precison + Recall} \tag{11}$$

2.3.2. Diversity measures

The diversity measure is used to measure the diversity degree of individual learners in one ensemble learning model. By comparing 10 diversity measures, Kuncheva and Whitaker [43] recommended the use of Q statistic based on ease of interpretation, moreover, the double failure (DF) was also suggested as a complement measure. Therefore, in the present study, Q statistic and double failure (DF) are employed for diversity measure. The lower the value of both, the more diversity there is between each pair of learners. The overall DF and Q statistic values of the model can be obtained by calculating the average of DF and Q-statistic values between each pair of learners.

Before introducing DF and Q statistic, it is necessary to introduce the following content: suppose there are n base learners, L_i and L_j $(i, j = 1, 2, \ldots, n, i \neq j)$ are two different learners, N^{11} (N^{00}) is the number of samples for which both learners L_i and L_j classify correctly (incorrectly), and N^{10} (N^{01}) represents the learner L_i (L_j) classifies correctly while the learner L_j (L_i) classifies incorrectly, as shown in Table 2.

The Q statistic between two learners L_i and L_j can be defined as Equation (12).

$$Q_{ij} = \frac{N^{11}N^{00} - N^{10}N^{01}}{N^{11}N^{00} + N^{10}N^{01}} \tag{12}$$

From Equation (12), it can be seen that if both learners classify correctly or incorrectly, indicating $N^{10} = N^{01} = 0$, then $Q_{ij} = 1$, in the case, the degree of diversity between L_i and L_j is the lowest; In contrast, if both learners obtain different results on the same sample, that is, $N^{11} = N^{00} = 0$, then $Q_{ij} = -1$, and the degree of diversity is the highest in this case.

The DF focuses on the sample that both learners L_i and L_j misclassify, which is defined as Equation (13).

$$DF_{ij} = \frac{N^{00}}{N} \tag{13}$$

If L_i and L_j always misclassify at the same time, the larger N^{00}, the larger DF_{ij}, and the lower the accuracy and diversity between the two learners.

2.4. Shapley Additive Explanations

In the practical application of machine learning models, it is not usually limited to improving the performance of the model developed, but exploring the reasons for the formation of the model's results, which could help to optimize the model's performance as well as to better understand the model itself.

Shapley Additive Explanations (SHAP) is a new method of model explanation that combines global and local explanations with the use of Shapley values in game theory [44]. For a subset of risk factors $S \subseteq F$ (where F represents the set of all the risk factors). Two models are trained to extract the impact of factor i. The first model $f_{S\cup\{i\}}(x_{S\cup\{i\}})$ is trained with factor i, while another model $f_S(x_S)$ is not trained with factor i. The difference in the output of the model $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$ is calculated for each possible subset by $S \subseteq F \setminus \{i\}$ [36], which is expressed as Equation (14).

$$\varnothing_i = \sum S \subseteq F \setminus \{i\} \frac{|S|!(|F| - |S| - 1)!}{|F|!} f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S) \tag{14}$$

3. Dataset and methodology application

3.1. Dataset

Official maritime accident investigation reports are widely recognized as a reliable source that can obtain objective and comprehensive information about the maritime accidents [45]. The present study collected 555 maritime accidents occurring from the range of 2011–2020 on the official website of 7 maritime investigation organizations (listed in Table 3). The collected maritime

Table 2
Combination of classification results of two learners.

	L_j correct	L_j incorrect
L_i correct	N^{11}	N^{10}
L_i incorrect	N^{01}	N^{00}

Table 3
Maritime investigation organizations reviewed in the study.

Maritime investigation organization	Abbreviation	Country
Maritime Safety Administration of People's Republic of China	MSA	China
Marine Accident Investigation Branch	MAIB	United Kingdom
United States National Transportation Safety Board	NTSB	United States
Australian Transport Safety Bureau	ATSB	Australia
Swedish Accident Investigation Board	SHK	Sweden
Korean Maritime Safety Tribunal	KMST	South Korea
Japan Transport Safety Board	JTSB	Japan

accident reports contain the five most frequently occurring types of maritime accident, namely collision, grounding, sinking, contact, and fire/explosion.

In order to ensure that the maritime accident reports involved seafarers' unsafe acts, this study initially screened the collected maritime accident reports and removed the maritime accidents that were entirely due to objective factors such as environmental factors and ship factors. Finally, 476 human-related maritime accident investigation reports were used for further analysis, and the fundamental statistical data of accident reports are displayed in Fig. 3.

In the study of Lan et al. [14], a theoretical framework for analyzing unsafe acts involved in maritime accidents was developed. With the help of the framework, this study explored the 476 accident reports and extracted 31 seafarers' unsafe acts and 44 risk factors. Then, Microsoft Excel 2013 software was used to construct a dataset including the identified factors. The descriptive statistic information of seafarers' unsafe acts and their risk factors in the dataset are shown in Table 4. For the 0/1 variable, Table 4 only shows the frequency when the variable is 1. It can be found that Failure to maintain proper lookout (U17) has the highest frequency (48.32 %) in seafarers' unsafe acts level, followed by Failure to determine the risk (U31) and Failure to take effective collision avoidance action early (U25). In organizational influence layer, Insufficient education and training (O3) has the highest frequency, reaching 30.88 %, followed by Poor competence (O5). Then, Inadequate safety management (S4) occurs most frequently (41.81 %) in unsafe supervision layer, and Lack of safety awareness (P10) occurs most frequently (13.66 %) in precondition for unsafe acts layer.

3.2. Individual learner development

The methodology proposed is conducted by R software using the "mlr3" package. Hyperparameter optimization of ML algorithms is essential for improving prediction accuracy and avoiding over fitting and under fitting problems. First, this study utilizes random search method combined with 10-fold cross-validation and 100 iterations for a large range of hyperparameter optimization search. According to the process: the dataset is split to 10 subsets, and then each subset is selected in turn as the validation set, and the other 9 subsets are employed as the training set. Within the specified range (the second column of Table 5), a set of hyperparameter is randomly selected to construct 10 models, and the performance of the hyperparameter sets is measured according to the model accuracy. The above process is iterated 100 times, and the hyperparameter set with the highest model accuracy is output, and the results are shown in the third column of Table 5. On this basis, the search range of hyperparameters is narrowed down, and the hyperparameters are refined and tuned using grid search and 10-fold cross-validation to obtain the final set of hyperparameters for individual model development, and the results are shown in the fourth column of Table 5.

The dataset is divided randomly into test set (20 %) and training set (80 %). The seven individual learners are developed with the optimal hyperparameters, and form the individual learner set L ($L = \{MLR, BP, SVM, KNN, CART, RF, XGBoost\}$). The model performance based on test set is provided in Fig. 4. It can be found that RF model achieves the highest predictive accuracy of 85.42 %, followed by the SVM model, with an accuracy of 82.29 %.

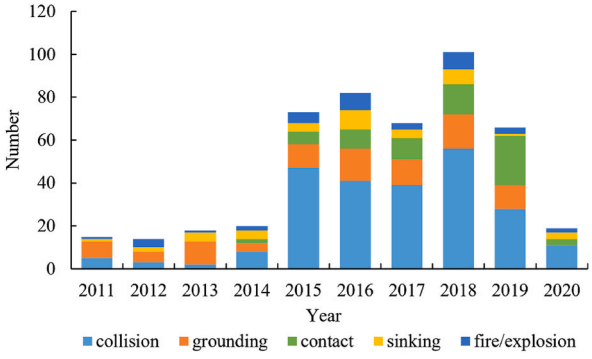


Fig. 3. Statistics of maritime accident reports examined in the study.

Table 4
Descriptive statistics of seafarers' unsafe acts and their risk factors.

Classification	Variable	Observation	Frequency
Organizational influence	Insufficient device (O1)	1-available; 0-unavailable	13.24 %
	Lack of operation certificate (O2)	1-available; 0- unavailable	2.73 %
	Insufficient education and training (O3)	1-available; 0- unavailable	30.88 %
	Insufficient manning (O4)	1-available; 0- unavailable	15.97 %
	Poor competence (O5)	1-available; 0- unavailable	23.53 %
	Poor information transmission of the company (O6)	1-available; 0- unavailable	1.05 %
	Lack of standardization (O7)	1-available; 0- unavailable	12.39 %
Unsafe supervision	Lack of route plan review (S1)	1-available; 0- unavailable	1.89 %
	Lack of supervision and guidance (S2)	1-available; 0- unavailable	29.41 %
	Navigation beyond the authorized areas (S3)	1-available; 0- unavailable	6.09 %
	Inadequate safety management (S4)	1-available; 0- unavailable	41.81 %
	Insufficient maintenance (S5)	1-available; 0- unavailable	11.13 %
	Inappropriate route plan (S6)	1-available; 0- unavailable	4.20 %
	Cargo defect (S7)	1-available; 0- unavailable	5.25 %
	Fail to correct the mistakes (S8)	1-available; 0- unavailable	2.73 %
	Ignore rules and regulations (S9)	1-available; 0- unavailable	4.41 %
Precondition for unsafe acts	Poor communication between ships (P1)	1-available; 0- unavailable	13.03 %
	Poor communication (ship-shore) (P2)	1-available; 0- unavailable	2.52 %
	Poor team communication (P3)	1-available; 0- unavailable	13.03 %
	Insufficient utilization of bridge resources (P4)	1-available; 0- unavailable	5.04 %
	Inadequate preparation (P5)	1-available; 0- unavailable	1.26 %
	Poor physical condition (P6)	1-available; 0- unavailable	0.42 %
	Poor emotional state (P7)	1-available; 0- unavailable	1.26 %
	Stress (P8)	1-available; 0- unavailable	0.21 %
	Alcohol/Drugs (P9)	1-available; 0- unavailable	1.05 %
	Lack of safety awareness (P10)	1-available; 0- unavailable	13.66 %
	Lack of situational awareness (P11)	1-available; 0- unavailable	5.46 %
	Distraction (P12)	1-available; 0- unavailable	5.88 %
	Fatigue (P13)	1-available; 0- unavailable	7.35 %
	Over-confidence (P14)	1-available; 0- unavailable	0.84 %
Environmental factors	Device failure (E1)	1-available; 0- unavailable	4.83 %
	Flooding (E2)	1-available; 0- unavailable	5.04 %
	Season (E3)	1-spring; 2-summer; 3-autumn; 4-winter	1:29.20 %; 2:20.38 %; 3:25.00 %; 4:25.41 %
	Time (E4)	1-day; 2-night	1:41.18 %; 2:58.82 %
	Ship type (E5)	1-cargo ship; 2-container; 3-tanker; 4-passenger ship; 5-fishing vessel; 6-others	1:45.17 %; 2:8.61 %; 3:9.45 %; 4:4.41 %; 5:9.03 %; 6:23.32 %
	Gross tonnage (E6)	1-≤1000t; 2-1001-5000t; 3-5001-10000t; 4->10000t	1:33.41 %; 2:32.53 %; 3:9.17 %; 4:24.89 %
	Ship length (E7)	1-≤100; 2->100	1:61.40 %; 2:38.60 %
	Ship age (E8)	1-≤10; 2->10	1:49.49 %; 2:50.51 %
	Complex navigation environment (E9)	1-available; 0- unavailable	22.48 %
	Busy traffic (E10)	1-available; 0- unavailable	22.06 %
	Poor visibility (E11)	1-available; 0- unavailable	16.81 %
	Strong wind and waves (E12)	1-available; 0- unavailable	7.56 %
	Tidal current effects (E13)	1-available; 0- unavailable	6.51 %
	External management (E14)	1-available; 0- unavailable	3.57 %
Seafarers' unsafe acts	Inadequate handover (U1)	1-available; 0- unavailable	1.68 %
	Failure to use protective equipment (U2)	1-available; 0- unavailable	2.94 %
	Failure to keep navigational equipment on working state (U3)	1-available; 0- unavailable	2.31 %
	Insufficient manning on bridge (U4)	1-available; 0- unavailable	8.82 %
	Failure to take safety measures in restricted visibility (U5)	1-available; 0- unavailable	0.42 %
	Failure to perform safety duties during berthing (U6)	1-available; 0- unavailable	3.36 %
	Drinking/Alcoholism (U7)	1-available; 0- unavailable	2.31 %
	OOW falls asleep on duty (U8)	1-available; 0- unavailable	4.20 %
	Violation of operational procedures (U9)	1-available; 0- unavailable	4.62 %
	Failure to check the course and position (OOW) (U10)	1-available; 0- unavailable	5.04 %
	Failure to check the planned route in time (OOW) (U11)	1-available; 0- unavailable	1.47 %
	Steering error (duty sailor) (U12)	1-available; 0- unavailable	0.42 %
	Insufficient use of navigational equipment (U13)	1-available; 0- unavailable	7.35 %

(*continued on next page*)

Table 4 (*continued*)

Classification	Variable	Observation	Frequency
	Over-reliance on navigational equipment (U14)	1-available; 0- unavailable	2.10 %
	Failure to exhibit proper light and shape (U15)	1-available; 0- unavailable	2.10 %
	Failure to make proper sound and light signals (U16)	1-available; 0- unavailable	9.24 %
	Failure to maintain proper lookout (U17)	1-available; 0- unavailable	48.32 %
	Failure to control the ship position (U18)	1-available; 0- unavailable	13.03 %
	Improper selection of anchoring position (U19)	1-available; 0- unavailable	1.68 %
	Improper emergency response measures (U20)	1-available; 0- unavailable	16.81 %
	Ignore alarm signals or warnings (U21)	1-available; 0- unavailable	0.42 %
	Failure of seafarers to follow best practices (U22)	1-available; 0- unavailable	1.26 %
	Failure to execute the planned route (U23)	1-available; 0- unavailable	1.89 %
	Unsafe speed (U24)	1-available; 0- unavailable	20.17 %
	Failure to take effective collision avoidance action early (U25)	1-available; 0- unavailable	27.94 %
	Failure to follow the rules in special waters such as narrow channel (U26)	1-available; 0- unavailable	17.86 %
	Failure to follow the rules in sight of one another (U27)	1-available; 0- unavailable	21.22 %
	Failure to follow the rules in restricted visibility (U28)	1-available; 0- unavailable	4.41 %
	Vigilance negligence (U29)	1-available; 0- unavailable	7.77 %
	Failure to determine the impact of environment on ship maneuvering (U30)	1-available; 0- unavailable	11.13 %
	Failure to determine the risk (U31)	1-available; 0- unavailable	31.09 %

Table 5
Hyperparameter optimization results.

Algorithm	Hyperparameters space	Random search	Grid search
MLR	Max. number of iterations: 1-200	21	21
BP	Max. number of iterations: 1–500; Number of neurons in the hidden layer: 5-20	46; 10	40; 10
SVM	Cost:0.1–10; gamma: 0–5; kernel: linear, polynomial, radial, sigmoid	0.73; 0.0253; radial	0.8; 0.023; radial
KNN	K:1-20	12	12
CART	Min. number of branch nodes: 1–20; Max. depth: 1-20	9; 16	10; 17
RF	Number of features: 2–10; Number of trees: 10–1000; Max. depth: 1-20	5; 608; 14	7; 596; 12
XGBoost	Max. depth: 1–8; gamma: 0–5; Min. child weight: 1–6; subsample: 0.5–1	8; 1.34; 1.15; 0.939	10; 1.3; 1.11; 0.8

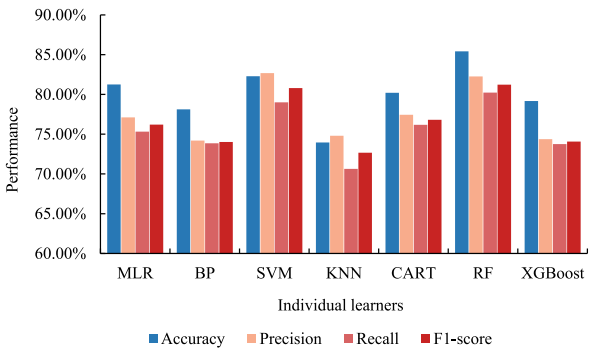


Fig. 4. The prediction performance of the individual learners.

3.3. Selective ensemble model development

Based on the two-stage selective ensemble learning method proposed in Section 2, the present study considers both model accuracy and diversity among the established individual learners, and selects some of them as base learners for Stacking ensemble learning. The base learners are selected using the following procedure:

S1: Remove the individual learners whose prediction accuracy does not exceed 75 %. The prediction accuracy of the KNN model is 73.96 %, which does not exceed 75 %, so it is removed from the set of individual learners L to avoid the negative impact of poorly

performing individual learners. As a result, $L = \{MLR, BP, SVM, CART, RF, XGBoost\}$;

S2: Select the initial base learner from the set of individual learners L. In Fig. 4, RF model provides the best prediction performance with a prediction accuracy of 85.42 % and an F1-score of 81.23 %, which has a good generalization ability. Therefore, the RF model is selected as the initial base learner E_1. As a result, $L = \{MLR, BP, SVM, CART, XGBoost\}, E = \{RF\}$;

S3: Select the second base learner from the set of individual learners L. The Q statistics and DF values between the remaining individual learners in L and E_1 are calculated separately, and the results are shown in the third row of Table 6. The Q statistics and DF values between the SVM model and the RF model are the smallest. Therefore, the SVM model is chosen as the second base learner E_2. As a result, $L = \{MLR, BP, CART, XGBoost\}, E = \{RF, SVM\}$;

S4: Select the third individual learner from the set of individual learners L. The Q statistics and DF values between the remaining individual learners in L and the Bagging-based ensemble model of E_1 and E_2 are calculated separately, and the results are shown in the fourth row of Table 6. The Q statistics and DF values between the XGBoost model and the Bagging-based ensemble model of RF and SVM are the smallest. Therefore, the XGBoost model is selected as the third base learner E_3. As a result, $L = \{MLR, BP, CART\}, E = \{RF, SVM, XGBoost\}$;

S5: The process is repeated until L is reorganized into an alternative sequence in E. The results of Q statistics and DF values during the process are shown in Table 6. Finally, $E = \{RF, SVM, XGBoost, MLR, CART, BP\}$;

S6: The top pre-defined individual learners that can optimize the prediction performance of the Stacking ensemble model are selected as the base learners. In the present study, the effects of individual learners on Stacking ensemble model are calculated separately by the order of individual learners in $E = \{RF, SVM, XGBoost, MLR, CART, BP\}$, and the results are illustrated in Fig. 4. The model performance reaches optimal when the first three individual learners (RF, SVM, and XGBoost) are selected for integration. Therefore, RF, SVM and XGBoost are selected as the base learners to develop the prediction model for maritime accident types.

Fig. 5, it can be found that the accuracy of the Stacking ensemble learning model is 87.50 % and the F1-score is 84.98 % when the first three individual learners (RF, SVM, and XGBoost) are selected for integration. Compared with the optimal individual learner RF, the accuracy of the selective ensemble learning model increased by 2.08 % and the F1-score improved by 3.75 %, which indicates that integrating multiple individual learners could enhance the prediction performance of the model. However, when all the individual learners are integrated, the prediction performance is not satisfactory, with an accuracy of 82.29 % and an F1-score of 76.11 %, which reveals the overfitting problem caused by too many individual learners, and the poor performance of some individual learners affects the reliability of the ensemble learning model. Therefore, in practical applications, Stacking ensemble learning model with the first three individual learners could help to improve the predictability of maritime accident types. The proposed two-stage selective ensemble learning method reduces the size of integration by eliminating redundant individual learners, so as to further improve the performance of the ensemble learning model.

3.4. Performance evaluation

To visually demonstrate the predictive performance of Stacking ensemble learning model developed, a comparative analysis is conducted with eight other prediction models, including three ensemble models and five single machine learning models. The five single machine learning models are MLR, SVM, BP, KNN, and CART, and the three ensemble models are RF, XGBoost, and the Bagging ensemble model of RF, SVM, and XGBoost. The comparison results of model performance are shown in Fig. 6.

Fig. 6 shows that among five single machine learning models, the SVM model has the best prediction performance (accuracy is 82.29 % and F1-score is 80.80 %), followed by the MLR model (accuracy is 81.25 % and F1-score is 76.21 %) and the CART model (accuracy is 80.21 % and F1-score is 76.81 %). The KNN model performs less well, with an accuracy of 73.96 %. For the three ensemble models, the RF model achieves the greatest predictive performance (accuracy is 85.42 % and F1-score is 81.23 %) and exceeds all the single machine learning models, which indicates that the ensemble models generally outperform single machine learning models. However, the prediction performance of XGBoost outperformed only the KNN and BP models, indicating that the overall integrated model may perform worse than single machine learning models, which is similar to the results of selective ensemble learning described above. The prediction model developed in this study reaches 87.50 % accuracy and 84.98 % F1-score, which is higher than the other models. The results reveal that the prediction model developed could obtain high accuracy and great generalization ability, which would be effectively used for maritime accident type prediction.

To better evaluate the model prediction performance, the result of confusion matrix is provided in Fig. 7. It can be seen that there is no misclassified sinking accident record, which means that the model could effectively predict sinking accidents. Fig. 8 also reflects

Table 6
Q statistics and DF values between the individual learners.

Model	Q statistics					DF values				
	SVM	XGBoost	MLR	CART	BP	SVM	XGBoost	MLR	CART	BP
RF model	0.9103	0.9677	0.9774	0.9507	0.9330	0.1250	0.1354	0.1458	0.1354	0.1345
Bagging-based model of RF and SVM		0.8780	0.9889	0.9633	0.9771		0.1146	0.1458	0.1354	0.1458
Bagging-based model of RF, SVM, XGB			0.8842	0.9103	0.9330			0.1146	0.1250	0.1345
Bagging-based model of RF, SVM, XGB, MLR				0.8919	0.9379				0.1250	0.1354
Bagging-based model of RF, SVM, XGB, MLR, CART					0.9346					0.1563

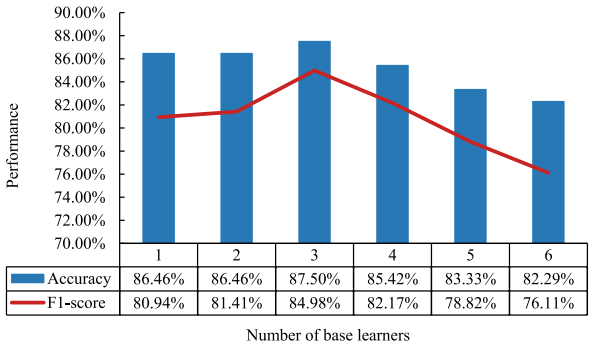


Fig. 5. The effect of individual learners on the stacking ensemble model.

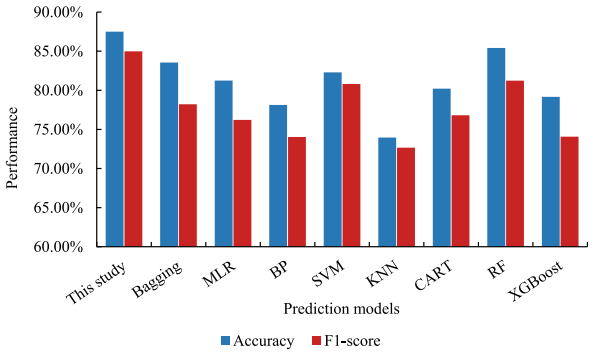


Fig. 6. The comparison results of model performance.

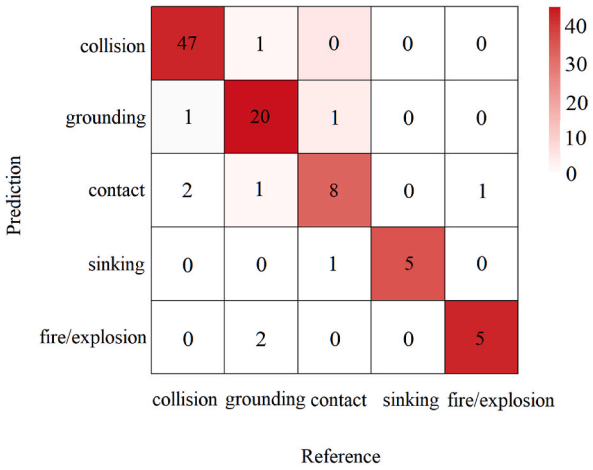


Fig. 7. Confusion matrix of maritime accident types.

similar result, the prediction precision of sinking accidents reaches 100 %, followed by collision accidents (94.00 %), grounding accidents (83.33 %) and fire/explosion accidents (83.33 %). However, the recall rate of contact accidents is relatively poor, with a recall of 66.67 %, indicating that some contact accidents are incorrectly predicted as other types. The confusion matrix shows that 4 contact accidents are misclassified into collision, grounding as well as fire/explosion accidents. It is worth noting that the dataset analyzed in the present study is collated manually, and any subjective issues may contribute to the misclassification.

3.5. Robustness test

After evaluating the model performance, the robustness should be examined. In the present study, the robustness of the nine models

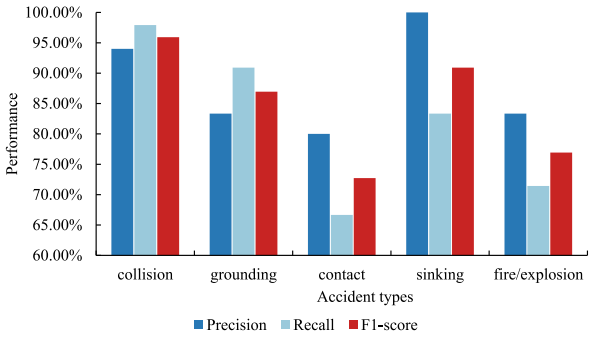


Fig. 8. Prediction performance by different type of maritime accident.

mentioned above is validated using the strategy proposed by Sarkar et al. [41]. The dataset is split into training set and test set by five different random seeds (1 %, 2 %, 3 %, 5 %, and 10 %), and then 10-fold cross-validation is measured for each of the five training sets. As shown in Fig. 9, the predictive performance of the model developed in the present study is relatively stable with high robustness under different conditions.

3.6. Model explanation analysis

To illustrate the contribution and importance of input features on various maritime accident types, the SHAP method is adopted to interpret and analyze the prediction model developed. The SHAP values are obtained by R software using the "shapper" package, the computational time is approximately 1320 s. Fig. 10 provides the ranking of feature importance, and the variables with lower feature importance are combined. Also, Fig. 10 clarifies the positive and negative effects of each input feature on the target variables. The scatter points in the figure indicate the different Shapley values of the feature variables, the color represents the high (yellow) and low (purple) values of the feature variables, and the density of the points indicates their distribution in the data set.

As shown in Fig. 10 (a), failure to maintain a proper lookout (U17) has the greatest effect on predicting collision accidents, and the higher the Shapley values, the higher the probability of collisions. Then, followed by failure to take effective collision avoidance action early (U25) and failure to follow the rules in sight of one another (U27), similarly, the probability of collision accidents increases with increasing feature values. These three features with great effects for predicting collision accidents are all belong to seafarers' unsafe acts, which also reflects that unsafe acts are the primary factors contributing to the occurrence of collisions. In contrast, environmental factors (strong wind and waves, and flooding) have negative impact on the occurrence of collisions, that is, the probability of collision accidents increases with decreasing feature values. In addition, ship type (E5) and ship age (E8) also have negative impact on the occurrence of collisions.

The results of the feature importance for grounding accidents are shown in Fig. 10 (b). The top five features that have significant effect on predicting grounding accidents are all belong to seafarers' unsafe acts. Among them, failure to maintain a proper lookout (U17), failure to take effective collision avoidance action early (U25) and violation of operational procedures (U9) have negative impact on the occurrence of grounding accidents, thus the probability of grounding accidents increases with decreasing feature values. Failure to check the course and position (OOW) (U10) and Insufficient use of navigational equipment (U13) have positive impact on the occurrence of grounding accidents, indicating that the higher the feature values, the higher the probability of grounding accident. Similar to collisions, strong wind and waves (E12) and flooding (E2) also have significant effect on predicting grounding accidents, however, strong wind and waves (E12) have positive impact on the occurrence of grounding accidents, that is, the probability of

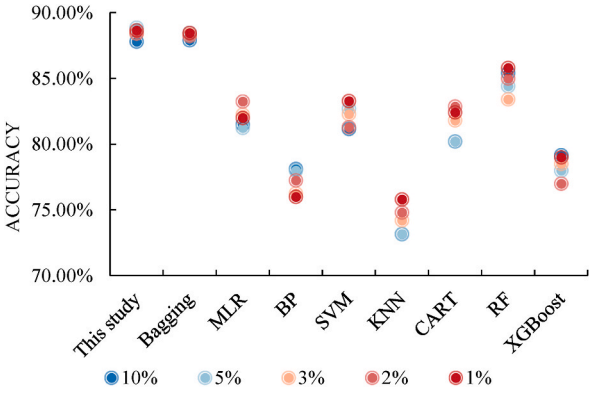


Fig. 9. Robustness test of the nine models.

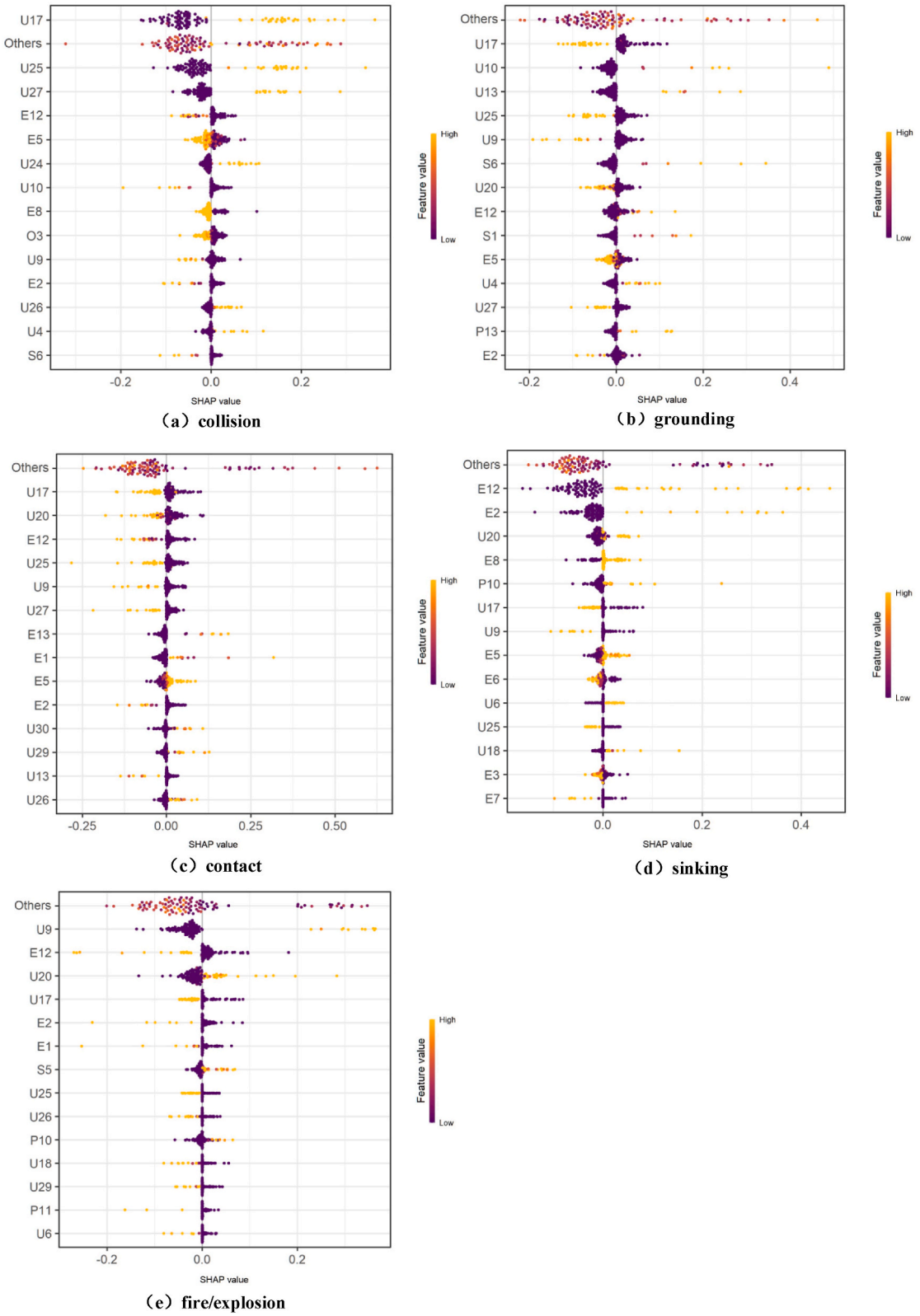


Fig. 10. Feature importance for different type of maritime accident.

grounding accidents increases with increasing feature values. At the unsafe supervision level, lack of route plan review (S1) and inappropriate route plan (S6) play a positive role in predicting the occurrence of grounding accidents, and when these two risk factors occur, the probability of causing grounding accidents is higher. Additionally, ship type (E5) has negative impact on grounding accidents, indicating that the probability of grounding accidents increases with decreasing feature values.

The results of the feature importance for contact accidents are shown in Fig. 10 (c), and the most important feature is still failure to maintain a proper lookout (U17), followed by failure to control the ship position (U20). However, they have negative effect on predicting the occurrence of contact accidents. In addition to seafarers' unsafe acts, environmental factors also play an important role in contact accidents. Among them, strong wind and waves (E12) and flooding (E2) have negative impact on the occurrence of contact accidents, indicating that the probability of contact accidents increases with decreasing feature values. Tidal current effect (E13) and device failure (E1) have positive impact on the occurrence of contact accidents, indicating that the probability of contact accidents increases with increasing feature values. In addition, ship type (E5) has positive impact on the occurrence of contact accidents.

The feature importance results for sinking accidents are shown in Fig. 10 (d). The top two important features are strong wind and waves (E12) and flooding (E2), both of which are environmental factors and have positive effect on predicting the occurrence of sinking accidents. Then, failure to control the ship position (U20) and lack of safety awareness (P10) play an important role in the occurrence of sinking accidents. In addition, ship type (E5) and ship age (E8) have positive impact on the occurrence of sinking accidents, that is, the probability of sinking accidents increases with increasing feature values. Meanwhile, gross tonnage (E6) and season (E3) have negative impact on the occurrence of sinking accidents, indicating that the probability of sinking accidents increases with decreasing feature values.

The feature importance results for fire/explosion accidents are shown in Fig. 10 (e). The most important feature is violation of operational procedures (U9), and the higher the feature values, the higher the probability of fire/explosion accidents. At the unsafe supervision level, insufficient maintenance (S5) has positive effect on predicting fire/explosion accidents, indicating that the probability of fire/explosion accidents increases with increasing feature values. In term of environmental factors, device failure (E1), flooding (E2), and strong wind and waves (E12) have negative effect on predicting fire/explosion accidents, the probability of fire/explosion accidents increases with decreasing feature values. In addition, lack of safety awareness (P10) and lack of situational awareness (P11) have important impact on predicting the occurrence of fire/explosion accidents.

4. Discussion

The prediction model of maritime accident type developed in the present study achieves great model performance for two main reasons. One of the reasons is that the model developed utilize the special structure of the Stacking model and the base learners integrated are based on heterogeneous machine learning algorithms. As a result, each algorithm has its own optimization criteria and classification strategy that allows describing the distribution patterns of the dataset in multiple perspectives. Aggregating individual learners constructed using these different algorithms may produce better predictive results than single machine learning models. On the other hand, conventional ensemble learning models aggregate all individual learners without being able to examined the quality of individual learners. Consequently, this study proposes a two-stage principle for model selection, which removes some poorly performing individual learners and selects only three high-variance and high-quality individual learners to form the first layer of the Stacking ensemble model, which is the essential reason why the model developed could obtain sufficiently good prediction results.

The performance benefits of the prediction model developed in the present study can also be reflected in the literature related to maritime safety. For instance, Lan et al. [46] developed a prediction model for the severity of maritime accidents with an accuracy of 80 % utilizing random forest, which is 7.5 % less accuracy than the prediction model developed in the present study, further highlighting the superiority of the two-stage selective ensemble learning method proposed. The results of model explanation analysis revealed that "unsafe acts" level played a greater role in predicting the type of maritime accidents compared to the other levels. However, Chen et al. [47] defined "precondition for unsafe acts" level as the primary contributor to the occurrence of maritime accidents (34.8 %), and hardware failure was the top risk factor in the level. In addition to differences in the target variables, one possible reason is that advances in science and technology have reduced hardware failure, while unsafe acts are more difficult to control due to the uncertainty. Additionally, some studies concluded that ship age [48] and ship type [48,49] could be associated with ship accident severity. Similarly, these factors are also identified as the factors that influencing the prediction of maritime accident types.

The results prove that the type of maritime accident occurred are not random, but have potential patterns and could be detected. Based on the results, some safety management recommendations are provided. (1) Shipping companies are suggested to establish the seafarers' unsafe act checklist for different type of maritime accident and arrange in descending order of importance. Based on this checklist, shipping companies could develop tailored safety education and training, and seafarers on board could conduct self-inspection and mutual supervision. (2) Shipping companies are suggested to formulate targeted education and training programs in accordance with the seafarers' unsafe act checklist and conduct regular navigation safety operation assessments. Meanwhile, based on the effect of ship type, gross tonnage, ship age and other ship factors on the prediction of maritime accidents, differentiated education and training contents are implemented for seafarers on different ships. For example, it is recommended to strengthen the training of professional collision avoidance knowledge for seafarers working on cargo ships and container ships with an age of less than ten years. (3) Shipping companies are advised to carry out regular and random inspections for the implementation of the safety management system on board, strengthen the daily supervision and management of the ship, and ensure long-term and effective supervision of the seafarers. In addition, safety management personnel on board should assume the responsibility of supervising and guiding seafarers in safe operations, ensuring that route plans are properly designed and reviewed, and correcting failures in a timely manner to minimize the occurrence of ship groundings. (4) With the support of computer vision and Internet of Things, it is recommended to develop a

monitoring and early warning platform based on the seafarers' unsafe act checklist, so as to realize real-time accurate monitoring of typical seafarers' unsafe acts. At the same time, the prediction model of maritime accident types developed in the present study also provides technical support for the realization of the monitoring and early warning of seafarers' unsafe acts to a certain extent.

Although the prediction model developed has many advantages, several limitations need to be highlighted. First, the dataset used in the present study is collated manually, and any subjective issues may affect the model performance. Meanwhile, this study only focuses on the five types of maritime accident since there is limited available data of unsafe acts involved in maritime accidents, which may affect the practicality of the prediction model in some extent. Therefore, it is necessary to collect and collate data carefully on other maritime accident types in future studies, and the approach of uncertainty quantification could adopt the method proposed by Abbaszadeh Shahri et al. [50]. Furthermore, this study lacks objective data on accidents such as flags, general location, etc. Future studies are suggested to update data and combine accident report data with AIS data, Lloyd's Register data, inspection reports, etc. may find useful results for predicting maritime accidents. With the amount of data increases, deep learning techniques, such as smart sustainable intelligent transportation systems [51], cognitive-radio-based internet of things networks [52], and visual perception and environment mapping algorithms [53] could be introduced for maritime accident prediction in further studies. Additionally, SHAP also has drawbacks, and other post-hoc explanation methods could be used for the same object. Therefore, future studies are suggested to compare the model explanation results obtained by different methods, and the problem about additivity constraints, explanations as contrastive statements, marginal contributions also should be discussed.

5. Conclusions

Without the prediction of risk factors at sea, maritime authorities could only make safety management measures passively to prevent maritime accidents. Even if traditional statistical methods are able to conduct risk prediction, it remains a challenge to ensure the accuracy and stability of the prediction model. To address this safety issue, the present study utilizes machine learning techniques to develop prediction model with explanations for maritime accident types. The purpose is to predict the types of maritime accident accurately and reliably, and provide useful references for maritime risk assessment and tailored preventions of different types of human-related maritime accidents, so as to improve the reliability of maritime operations. The main contributions of the present study are as follows.

(1) The present study proposes a two-stage selective ensemble learning method to develop the prediction model of maritime accident types. The model developed achieves good prediction performance with an accuracy of 87.50 % and an F1-score of 84.98 %. Compared with the general ensemble model with all the individual learners, the proposed method eliminates the redundant individual learners, avoids the over fitting problem, and improves prediction performance. Additionally, through the performance comparison and robustness test with eight common machine learning models, the prediction model developed shows better accuracy and stability.
(2) The present study integrated SHAP method with selective ensemble learning to make the prediction model with explanation, which could clarify the effective mapping association and the association degree between the model prediction results and seafarers' unsafe acts.

The present study adopts unsafe acts of seafarers and their risk factors to predict the types of maritime accident, and introduces selective ensemble learning method to reduce the possibility of single machine learning models misclassifying and falling into local optimal solutions, so as to improve the prediction accuracy and stability of the model developed, which provides a new way to evaluate maritime safety quickly, accurately and reliably. Moreover, the model developed is explainable due to SHAP method, indicating that the model is able to identify seafarers' unsafe acts and their risk factors playing significant role in predicting the types of maritime accident, which helps stakeholders involved to develop tailored and proactive countermeasures to prevent the occurrence of maritime accidents. In addition, as the methodology is unbiased, it could be used in other industries in which human factors serve a significant role in accidents. Coal mines, railroads and aviation are possible candidates to employ the methodology proposed.

Funding

This study is supported by the National Key Research and Development Program of China (grand No. 2019YFB1600602).

Data availability

Data will be made available on reasonable request.

CRediT authorship contribution statement

He Lan: Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Shutian Wang:** Conceptualization, Resources, Validation, Visualization, Writing – review & editing. **Wenfeng Zhang:** Conceptualization, Supervision, Validation.

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.

References

[1] X. Zhou, Spatial risk assessment of maritime transportation in offshore waters of China using machine learning and geospatial big data, Ocean Coast Manag. 247 (2024), https://doi.org/10.1016/j.ocecoaman.2023.106934.

[2] L. Ma, X. Ma, Y. Liu, W. Deng, H. Lan, Risk assessment of coupling links in hazardous chemicals maritime transportation system, J. Loss Prev. Process. Ind. 82 (2023) 105011, https://doi.org/10.1016/j.jlp.2023.105011.

[3] MSA. Accident investigation reports. Available online:https://www.msa.gov.cn/html/hxaq/sgjx/index.html. (accessed on April 5, 2021).

[4] Z. He, C. Wang, J. Gao, Y. Xie, Assessment of global shipping risk caused by maritime piracy, Heliyon 9 (2023) e20988, https://doi.org/10.1016/j.heliyon.2023.e20988.

[5] W. Qiao, Y. Liu, X. Ma, Y. Liu, Human factors analysis for maritime accidents based on a dynamic fuzzy bayesian network, Risk Anal. 40 (2020) 957–980, https://doi.org/10.1111/risa.13444.

[6] A. Coraddu, L. Oneto, B. Navas de Maya, R. Kurt, Determining the most influential human factors in maritime accidents: a data-driven approach, Ocean Eng. 211 (2020), https://doi.org/10.1016/j.oceaneng.2020.107588.

[7] O. Soner, U. Asan, M. Celik, Use of HFACS–FCM in fire prevention modelling on board ships, Saf. Sci. 77 (2015) 25–41, https://doi.org/10.1016/j.ssci.2015.03.007.

[8] Y. Wang, S. Fu, Framework for process analysis of maritime accidents caused by the unsafe acts of seafarers: a case study of ship collision, J. Mar. Sci. Eng. 10 (2022), https://doi.org/10.3390/jmse10111793.

[9] J. Yang, G. Ye, Q. Xiang, M. Kim, Q. Liu, H. Yue, Insights into the mechanism of construction workers' unsafe behaviors from an individual perspective, Saf. Sci. 133 (2021), https://doi.org/10.1016/j.ssci.2020.105004.

[10] S. Fan, Z. Yang, Accident data-driven human fatigue analysis in maritime transport using machine learning, Reliab. Eng. Syst. Saf. 241 (2024), https://doi.org/10.1016/j.ress.2023.109675.

[11] K. Wróbel, Searching for the origins of the myth: 80% human error impact on maritime safety, Reliab. Eng. Syst. Saf. 216 (2021) 107942, https://doi.org/10.1016/j.ress.2021.107942.

[12] B. Wu, T.L. Yip, X. Yan, C. Guedes Soares, Review of techniques and challenges of human and organizational factors analysis in maritime transportation, Reliab. Eng. Syst. Saf. 219 (2022) 108249, https://doi.org/10.1016/j.ress.2021.108249.

[13] S. Yildiz, Ö. Uğurlu, J. Wang, S. Loughney, Application of the HFACS-PV approach for identification of human and organizational factors (HOFs) influencing marine accidents, Reliab. Eng. Syst. Saf. 208 (2021) 107395, https://doi.org/10.1016/j.ress.2020.107395.

[14] H. Lan, X. Ma, W. Qiao, L. Ma, On the causation of seafarers' unsafe acts using grounded theory and association rule, Reliab. Eng. Syst. Saf. 223 (2022) 108498, https://doi.org/10.1016/j.ress.2022.108498.

[15] U. Yıldırım, E. Başar, Ö. Uğurlu, Assessment of collisions and grounding accidents with human factors analysis and classification system (HFACS) and statistical methods, Saf. Sci. 119 (2019) 412–425, https://doi.org/10.1016/j.ssci.2017.09.022.

[16] C. Chauvin, S. Lardjane, G. Morel, J.-P. Clostermann, B. Langard, Human and organisational factors in maritime accidents: analysis of collisions at sea using the HFACS, Accid. Anal. Prev. 59 (2013) 26–37, https://doi.org/10.1016/j.aap.2013.05.006.

[17] A. Graziano, A.P. Teixeira, C. Guedes Soares, Classification of human errors in grounding and collision accidents using the TRACEr taxonomy, Saf. Sci. 86 (2016) 245–257, https://doi.org/10.1016/j.ssci.2016.02.026.

[18] G. Terti, I. Ruin, J.J. Gourley, P. Kirstetter, Z. Flamig, J. Blanchet, A. Arthur, S. Anquetin, Toward probabilistic prediction of flash flood human impacts, Risk Anal. 39 (2019) 140–161.

[19] R. Zhu, X. Hu, J. Hou, X. Li, Application of machine learning techniques for predicting the consequences of construction accidents in China, Process Saf. Environ. Protect. 145 (2021) 293–302, https://doi.org/10.1016/j.psep.2020.08.006.

[20] S. Sarkar, S. Vinay, R. Raj, J. Maiti, P. Mitra, Application of optimized machine learning techniques for prediction of occupational accidents, Comput. Oper. Res. 106 (2019) 210–224, https://doi.org/10.1016/j.cor.2018.02.021.

[21] X. Zhang, S. Mahadevan, Ensemble machine learning models for aviation incident risk prediction, Decis. Support Syst. 116 (2019) 48–63, https://doi.org/10.1016/j.dss.2018.10.009.

[22] 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.

[23] Z. Ma, G. Mei, S. Cuomo, An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors, Accid. Anal. Prev. 160 (2021) 106322, https://doi.org/10.1016/j.aap.2021.106322.

[24] A. Rawson, M. Brito, A survey of the opportunities and challenges of supervised machine learning in maritime risk analysis, Transport Rev. 43 (2022) 108–130, https://doi.org/10.1080/01441647.2022.2036864.

[25] R.J. Bye, A.L. Aalberg, Maritime navigation accidents and risk indicators: an exploratory statistical analysis using AIS data and accident reports, Reliab. Eng. Syst. Saf. 176 (2018) 174–186, https://doi.org/10.1016/j.ress.2018.03.033.

[26] Z. Jiang, L. Zhang, W. Li, A machine vision method for the evaluation of ship-to-ship collision risk, Heliyon 10 (2024) e25105, https://doi.org/10.1016/j.heliyon.2024.e25105.

[27] A. Rawson, M. Brito, Z. Sabeur, L. Tran-Thanh, A machine learning approach for monitoring ship safety in extreme weather events, Saf. Sci. 141 (2021) 105336, https://doi.org/10.1016/j.ssci.2021.105336.

[28] S. Kaisar, A. Chowdhury, Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests, ICT Express 8 (2022) 563–568, https://doi.org/10.1016/j.icte.2022.02.011.

[29] H. Lv, K. Yan, Y. Guo, Q. Zou, A.E. Hesham, B. Liu, AMPpred-EL: an effective antimicrobial peptide prediction model based on ensemble learning, Comput. Biol. Med. 146 (2022) 105577, https://doi.org/10.1016/j.compbiomed.2022.105577.

[30] Z. Wang, H. Wen, Y. Su, W. Shen, J. Ren, Y. Ma, et al., Insights into ensemble learning-based data-driven model for safety-related property of chemical substances, Chem. Eng. Sci. 248 (2022) 117219, https://doi.org/10.1016/j.ces.2021.117219.

[31] J. Sun, S. Wu, H. Zhang, X. Zhang, T. Wang, Based on multi-algorithm hybrid method to predict the slope safety factor– stacking ensemble learning with bayesian optimization, J. Comput. Sci. 59 (2022) 101587, https://doi.org/10.1016/j.jocs.2022.101587.

[32] H. Zhang, S. Wu, X. Zhang, L. Han, Z. Zhang, Slope stability prediction method based on the margin distance minimization selective ensemble, Catena 212 (2022) 106055, https://doi.org/10.1016/j.catena.2022.106055.

[33] Z.H. Zhou, J. Wu, W. Tang, Ensembling neural networks: many could be better than all, Artif. Intell. 137 (2002) 239–263, https://doi.org/10.1016/S0004-3702(02)00190-X.

[34] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38, https://doi.org/10.1016/j.artint.2018.07.007.

[35] R. Xu, F. Luo, Risk prediction and early warning for air traffic controllers' unsafe acts using association rule mining and random forest, Saf. Sci. 135 (2021) 105125, https://doi.org/10.1016/j.ssci.2020.105125.

[36] X. Wen, Y. Xie, L. Jiang, Y. Li, T. Ge, On the interpretability of machine learning methods in crash frequency modeling and crash modification factor development, Accid. Anal. Prev. 168 (2022) 106617, https://doi.org/10.1016/j.aap.2022.106617.

[37] Y. Kim, Y. Kim, Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models, Sustain. Cities Soc. 79 (2022) 103677, https://doi.org/10.1016/j.scs.2022.103677.
[38] C. Yang, M. Chen, Q. Yuan, The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: an exploratory analysis, Accid. Anal. Prev. 158 (2021) 106153, https://doi.org/10.1016/j.aap.2021.106153.
[39] Y.M. Goh, C.U. Ubeynarayana, K.L.X. Wong, B.H.W. Guo, Factors influencing unsafe behaviors: a supervised learning approach, Accid. Anal. Prev. 118 (2018) 77–85, https://doi.org/10.1016/j.aap.2018.06.002.
[40] P. Bhattacharjee, V. Dey, U.K. Mandal, S. Paul, Quantitative risk assessment of submersible pump components using interval number-based multinomial logistic regression model, Reliab. Eng. Syst. Saf. 226 (2022) 108703, https://doi.org/10.1016/j.ress.2022.108703.
[41] S. Sarkar, A. Pramanik, J. Maiti, G. Reniers, Predicting and analyzing injury severity: a machine learning-based approach using class-imbalanced proactive and reactive data, Saf. Sci. 125 (2020) 104616, https://doi.org/10.1016/j.ssci.2020.104616.
[42] K. Koc, Ö. Ekmekcioğlu, A.P. Gurgun, Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers, Autom. ConStruct. 131 (2021) 103896, https://doi.org/10.1016/j.autcon.2021.103896.
[43] L.I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Mach. Learn. 51 (2003) 181–207.
[44] I.U. Ekanayake, D.P.P. Meddage, U. Rathnayake, A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP), Case Stud. Constr. Mater. 16 (2022) e01059, https://doi.org/10.1016/j.cscm.2022.e01059.
[45] X. Ma, H. Lan, W. Qiao, B. Han, H. He, On the causation correlation of maritime accidents based on data mining techniques, Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability (2022) 1–15, https://doi.org/10.1177/1748006x221131717.
[46] H. Lan, X. Ma, W. Qiao, W. Deng, Determining the critical risk factors for predicting the severity of ship collision accidents using a data-driven approach, Reliab. Eng. Syst. Saf. 230 (2023) 108934, https://doi.org/10.1016/j.ress.2022.108934.
[47] S.-T. Chen, A. Wall, P. Davies, Z. Yang, J. Wang, Y.-H. Chou, A Human and Organisational Factors (HOFs) analysis method for marine casualties using HFACS-Maritime Accidents (HFACS-MA), Saf. Sci. 60 (2013) 105–114, https://doi.org/10.1016/j.ssci.2013.06.009.
[48] L. Wang, Z. Yang, Bayesian network modelling and analysis of accident severity in waterborne transportation: a case study in China, Reliab. Eng. Syst. Saf. 180 (2018) 277–289, https://doi.org/10.1016/j.ress.2018.07.021.
[49] H. Wang, Z. Liu, X. Wang, T. Graham, J. Wang, An analysis of factors affecting the severity of marine accidents, Reliab. Eng. Syst. Saf. 210 (2021) 107513, https://doi.org/10.1016/j.ress.2021.107513.
[50] A. Abbaszadeh Shahri, C. Shan, S. Larsson, A novel approach to uncertainty quantification in groundwater table modeling by automated predictive deep learning, Nat. Resour. Res. 31 (2022) 1351–1373, https://doi.org/10.1007/s11053-022-10051-w.
[51] A. Novak, A.N. Sedlackova, M. Vochozka, H.P. Gheorghe, Big data-driven governance of smart sustainable intelligent transportation systems: autonomous driving behaviors, predictive modeling techniques, and sensing and computing technologies, Contemp. Read. Law Soc. Justice 14 (2022) 100–117, https://doi.org/10.22381/CRLSJ14220226.
[52] X. Fernando, G. Lăzăroiu, Spectrum sensing, clustering algorithms, and energy-harvesting technology for cognitive-radio-based internet-of-things networks, Sensors 23 (2023), https://doi.org/10.3390/s23187792.
[53] M. Andronie, G. Lăzăroiu, O.L. Karabolevski, R. Ștefănescu, I. Hurloiu, A. Dijmărescu, et al., Remote big data management tools, sensing and computing technologies, and visual perception and environment mapping algorithms in the internet of robotic things, Electronics 12 (2022), https://doi.org/10.3390/electronics12010022.