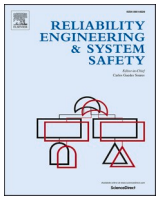
Reliability Engineering and System Safety 266 (2026) 111648

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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



Improving maritime accident severity prediction accuracy: A holistic machine learning framework with data balancing and explainability techniques

Wenjie Cao [a,b], Xinjian Wang [a,b,c,d,*], Yuanjun Feng [e], Jingen Zhou [g], Zaili Yang [b,f,**]

[a] Navigation College, Dalian Maritime University, Dalian 116026, PR China
[b] Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool L3 3AF, UK
[c] Key Laboratory of Navigation Safety Guarantee of Liaoning Province, Dalian 116026, PR China
[d] State Key Laboratory of Maritime Technology and Safety, Dalian 116026, PR China
[e] School of Management, University of Liverpool, Liverpool L69 7ZH, UK
[f] Transport Engineering College, Dalian Maritime University, Dalian 116026, PR China
[g] Department of Operations Management, Supply Chain and Information Systems, KEDGE Business School, Marseille 13288, France

ARTICLE INFO

Keywords:
Maritime safety
Marine accidents
Machine learning
Explainable AI
SHAP
LIME

ABSTRACT

Accurately predicting the severity of maritime accidents is crucial for enhancing safety management and minimizing operational risks. Traditional prediction models, however, often suffer from the challenges resulted from unbalanced datasets and the complexity of multidimensional factors. This study aims to develop an integrated prediction framework incorporating six data balancing techniques to effectively address category imbalance and enhance model predictive robustness. Additionally, eight well-established machine learning models are utilized, with their performance optimized through hyperparameter tuning and cross-validation. To interpret the model results, SHapley Additive exPlanations (SHAP) are applied for global feature contribution analysis, while Local Interpretable Model-agnostic Explanations (LIME) provide local interpretations, enabling an in-depth understanding of feature-specific impacts on predictions. The results indicate that the combination of RandomOverSampler and CatBoost achieves optimal performance across all metrics, with an accuracy of 86.45%, precision of 84.38%, recall of 89.70%, F1-score of 86.81%, and ROC AUC of 93.69%. The analysis identifies accident type, ship type, engine power and gross tonnage as the key features influencing accident severity prediction. Furthermore, the integrated explanatory framework combining SHAP and LIME elucidates both the individual contributions and the collective impact of these features, along with the direction and magnitude of their influence on individual predictions, ensuring model transparency and interpretability. This study advances the prediction of maritime accident severity and provides a robust scientific basis for decision-making in maritime safety, enabling policymakers and industrial stakeholders to make accurate and reliable risk-informed decisions. **The source code is publicly available at**: https://github.com/AdvMarTech/BalancedMaritimeAccidentXAI.

1. Introduction

With the rapid development and growing complexity of the global shipping industry, the rising concerns on severity of maritime accidents present significant challenges to shipping safety, environmental protection, and economic stability [1–3]. According to the European Maritime Safety Agency (EMSA), 26,595 maritime accidents were reported between 2014 and 2023, averaging 2,660 accidents per year [4]. The overall number of maritime accidents remains high, with severe accidents consistently accounting for a significant proportion over the past decade. For instance, by mid-2023, serious and very serious accidents accounted for 29.5% of all reported accidents [4]. This ongoing

* Corresponding author at: Navigation College, Dalian Maritime University, Dalian 116026, PR China.
** Corresponding author at: Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool L3 3AF, UK.
E-mail addresses: wangxinjian@dlmu.edu.cn (X. Wang), Z.Yang@ljmu.ac.uk (Z. Yang).

https://doi.org/10.1016/j.ress.2025.111648
Received 24 July 2025; Received in revised form 14 August 2025; Accepted 31 August 2025
Available online 1 September 2025

trend underscores the substantial challenges that maritime accident severity poses to shipping safety management [5,6].

The severity of maritime accidents, a critical indicator for evaluating their impact, directly influences shipping safety, the optimization of emergency response measures, and the enhancement of risk management systems, while also exerting a profound influence on the sustainable development of the global shipping industry [7,8]. With the globalization of shipping activities and the increasing size of vessels, predicting, controlling and preventing maritime accidents have become progressively more challenging [9,10]. Consequently, enhancing the accuracy of accident severity prediction has emerged as a critical research agenda item for ensuring shipping safety, optimizing emergency response measures, and strengthening risk management systems [5].

The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) techniques has led to their increasing adoption in maritime accident prediction, particularly for forecasting accident severity [2,7,10]. With its robust pattern recognition capabilities and adaptive optimization features, ML effectively manages complex, multidimensional, and highly nonlinear relationships within maritime accident data, particularly in environments with multiple intertwining factors, demonstrating significant predictive potential [11]. However, maritime accident data often suffer from severe category imbalance, which markedly undermines the performance of traditional ML models in predicting minority class samples [12]. To address this challenge, this study aims to develop a holistic framework embracing new supportive approaches to improving the effectiveness of maritime accident severity prediction by integrating multiple advanced data balancing techniques with ML models. Specifically, it seeks to mitigate the data imbalance issue by employing methods such as Synthetic Minority Over-sampling Technique (SMOTE) and KMeans Synthetic Minority Over-sampling Technique (KMeansSMOTE) [13], thereby enhancing the model's predictive capability and robustness. Additionally, by applying interpretable techniques like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), the model's transparency is increased, enabling decision-makers in maritime safety by improved scientific and informed predictions.

Additionally, as integrated learning algorithms and interpretability techniques continue to advance, maritime accident severity prediction has evolved into a multifaceted problem requiring consideration from multiple levels and perspectives. The challenge now lies in enhancing model interpretability while maintaining prediction accuracy, and in combining various data balancing techniques, ML algorithms, and interpretability methods to provide practical support for maritime accident prevention and risk control. This has become a central issue in the field of maritime safety, both now and in the future. Therefore, this study will address the following main questions:

1) Which ML models exhibit the best performance in predicting maritime accident severity?

2) Which data balancing techniques are the most effective in addressing the issue of unbalanced data in maritime accident severity prediction, and how can they enhance the predictive performance of the models?

3) What are the key features that significantly influence maritime accident severity, and how can their contributions be quantified and interpreted?

4) How can the combination of global feature contribution analysis (e.g. SHAP) and local interpretation methods (e.g. LIME) help improve the model's transparency and its practical applicability?

By systematically addressing these issues, this study makes contributions by providing a new theoretical foundation and practical guidance for maritime accident severity prediction, thereby offering scientific support for maritime safety management, risk decision-making, and emergency response.

The rest of the paper is structured as follows: Section 2 reviews the relevant literature, analyses the shortcomings and challenges of existing studies, and defines the state of the art in the field of this study. Section 3 introduces the research methodology, covering dataset construction, data balancing techniques, ML model selection and optimization, and the application of interpretable methods. Section 4 presents the experimental results, compares the performance of different methods, and analyses the model from both overall and local perspectives using interpretable techniques. This section also discusses the value and practical significance of each technique to prediction accuracy. Section 5 provides theoretical insights and practical guidance based on the experimental results, highlighting the study's academic contributions and its implications for real-world maritime safety management. Section 6 summarizes the entire paper, presents the research conclusions, discusses the limitations, and outlines the future research directions.

2. Literature review

This section systematically reviews state-of-the-art methodologies for maritime accident prediction and risk analysis. The review is structured into distinct themes to address the multifaceted challenges in maritime safety management. Firstly, it traces the evolution of maritime accident prediction techniques, from early rule-based expert systems and traditional statistical models to modern ML approaches, highlighting both advancements and limitations in handling complex, high-dimensional, and non-linear data. Secondly, it examines the role of data balancing techniques in risk analysis, emphasizing their importance in mitigating data imbalance, a factor that significantly impacts the performance of classification algorithms across various domains. Finally, it explores the integration of explainable artificial intelligence (XAI) in high-risk decision-making, focusing on its ability to enhance transparency and interpretability by addressing the "black-box" nature of ML models. This structured review not only provides a comprehensive foundation for understanding current methodologies but also establishes the basis for the novel contributions of this study.

2.1. Progress and challenges in maritime accident prediction

Over the past decade, maritime accident prediction methods have evolved from traditional statistical models to modern ML techniques, each demonstrating unique strengths and limitations in various contexts [2,7,10]. Ma et al. [14] and Kandel and Baroud [15] exemplified this transition through large-scale, data-driven maritime studies that foreground predictive modelling and factor effects at national and regional scales, respectively.

Early maritime accident prediction primarily relied on rule-based expert systems and statistical methods. Ung [16] proposed a predictive system grounded in expert knowledge and empirical rules, offering high interpretability by evaluating maritime accidents using predefined criteria. However, Akyuz et al. [17] highlighted that the empirical rule-based approach is constrained in its predictive capability due to its inability to account for the complex, non-linear relationships inherent in the maritime environment. With advancements in data mining techniques, Bayesian methods have gained widespread application in maritime accident prediction, particularly for modelling accident probabilities and managing uncertainty [18,19]. Although not traditionally classified as a ML approach, Bayesian methods have become integral to modern ML. Cao et al. [12] developed a data-driven Bayesian network (BN)-based probabilistic model for accidents, which effectively modelled conditional dependencies between variables and excelled in handling uncertainty and missing data. Despite achieving notable success in certain applications, the Bayesian approach faces limitations in addressing high-dimensional data and complex non-linear relationships. Kong et al. [20] and Li et al. [21] further advanced BN-based maritime safety assessment and collision-risk analysis by embedding scenario analysis and introducing a broader global perspective, thereby improving assessment accuracy under data-sparse conditions.

In recent years, ML methods, particularly ensemble learning

algorithms, have made significant advancements in maritime accident prediction. Yang et al. [22] analysed maritime accident data using Random Forest (RF) and Gradient Boosting Machine (GBM), demonstrating that these methods, through iterative optimization, can automatically identify potential patterns in the data and deliver improved predictive performance in complex maritime environments. Notably, in the context of high-dimensional data and intricate interactions, ML methods excel at uncovering implicit features that are challenging for traditional statistical approaches to detect. ML methods demonstrate such advantages not only in predicting accident occurrence but also in forecasting accident severity. For example, Wang et al. [11] compared various approaches for accident severity prediction, including classical statistical models such as the autoregressive integrated moving average (ARIMA) and ML techniques like back-propagation neural networks (BPNN) and support vector regression (SVR). Their findings indicate that while ARIMA is a robust statistical model capable of classifying maritime accidents in high-dimensional feature spaces and capturing certain non-linear relationships, ML approaches offer additional advantages. Specifically, ML models can better capture complex interactions and achieve higher predictive accuracy, particularly in scenarios with limited sample sizes. Furthermore, Kim and Lim [23] applied the K-Nearest Neighbours (KNN) algorithm to predict the severity of maritime accidents. Their study demonstrated that the KNN method accurately classifies accidents in multidimensional feature spaces by measuring sample similarities. This approach is particularly effective for datasets characterized by significant similarities or distinct local structural features. The further advancement of ML methods hinges on their integration with other techniques, particularly through hybrid approaches and multi-model combination frameworks. These frameworks address the limitations of single models when analysing complex maritime data by leveraging the strengths of multiple models. Atak and Arslanoglu [24] proposed a framework that combines ML with Bayesian inference, utilizing weighted averaging to integrate prediction results and enhance both accuracy and stability. This multi-model fusion framework exhibited strong generalization capabilities in maritime accident prediction.

ML methods have shown substantial advantages in predicting the severity of maritime accidents, particularly in addressing complex nonlinear relationships and high-dimensional data. By automatically identifying patterns within the data, ML effectively captures intricate interactions between accident severity and various factors, enhancing both the accuracy and stability of predictions. However, relatively few studies have specifically applied ML techniques to maritime accident severity prediction. While ML has proven effective in handling nonlinear relationships and high-dimensional challenges in related fields, its application in this domain remains limited. Although showing some attractiveness, using ML to predict maritime accident severity still reveals some practical challenges, including the quality of the data, choice of ML in terms of prediction performance under different scenarios and interpretation of the predicted results through ML approaches. Therefore, this study seeks to establish itself as a pioneering effort to address these challenges through a holistic approach, ultimately delivering a practical and robust framework to advancing maritime accident prediction capabilities.

2.2. Application of data balancing techniques in risk analysis

The widespread application of internet information technology across various industries has led to rapid growth in data volume and increasing complexity in data types. A significant portion of the data collected exhibits clear imbalance characteristics, with substantial variations in the quantity and proportion of different data categories [25]. Data imbalance is a prevalent issue in risk analysis, particularly in areas such as financial fraud detection, medical diagnosis, and traffic accident analysis [26]. In these fields, the dominance of data from the majority category often masks the sample information from the minority category, resulting in a higher misclassification rate for the minority class. This bias, when combined with accuracy or loss functions as the optimization objective, compromises the predictive performance of classification algorithms. The data imbalance problem is also significant in maritime accident severity prediction [2]. To address this issue, data balancing techniques offer notable advantages, effectively enhancing the model's ability to identify rare event categories and improving both prediction performance and result robustness.

To mitigate the impact of data imbalance on classification performance, researchers have proposed various solutions from both the algorithmic and data perspectives. Algorithmic-level solutions address this issue by modifying the model's learning strategy to enhance its ability to recognize samples from underrepresented classes. In the realm of data synthesis methods, Chawla et al. [27] introduced the SMOTE, an algorithm that balances datasets by interpolating between minority samples to generate synthetic data. This approach significantly improves the model's ability to classify minority samples. Building on SMOTE, Han et al. [28] developed Borderline Synthetic Minority Over-sampling Technique (BorderlineSMOTE), which focuses on oversampling minority samples near the decision boundary. This enhancement strengthens the model's classification performance in complex boundary regions. Within reliability research, Wu et al. [29] proposed a model-agnostic framework for class-imbalanced learning, Ding et al. [30] developed deep imbalanced domain adaptation that preserved minority-class structure under distribution shift, and Wu et al. [31] introduced a semi-supervised approach that simultaneously addressed imbalance and out-of-distribution samples.

Despite the success of various data balancing approaches across different domains, practical applications often necessitate selecting methods tailored to specific tasks. Consequently, researchers have shifted their focus on combining and comparing multiple data balancing techniques, emphasizing that a single method often fails to consistently achieve optimal results in all scenarios. Abdulsadig and Rodriguez-Villegas [32] employed hybrid sampling methods, including Synthetic Minority Over-sampling Technique-Tomek Links (SMOTETomek) and Synthetic Minority Over-sampling Technique - Edited Nearest Neighbours (SMOTEENN), to enhance the representativeness of minority classes by combining over-sampling and under-sampling techniques. Their approach demonstrated improved performance in complex and noisy high-risk accident prediction tasks. In recent years, studies have investigated the combined impact of integrating diverse data balancing strategies across multiple fields. Korkmaz et al. [33] integrated SMOTE, Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN), and BorderlineSMOTE with Artificial Neural Networks (ANNs), achieving improved accuracy in firewall packet classification, particularly in identifying minority class samples. This demonstrated the significant advantages of leveraging the synergistic effects of multiple data balancing techniques. Similarly, Abou Elassad et al. [34] combined SMOTE and ADASYN with various ML models to enhance classification performance in traffic accident prediction. Their approach effectively addressed class imbalance by augmenting the weights of minority class samples, thereby improving the model's capability to predict accidents with limited data.

The application of data balancing techniques has proven effective in mitigating the adverse effects of data imbalance on classification model performance, particularly in identifying high-risk accidents. Comparative analyses of multiple balancing strategies have demonstrated significant advantages by optimizing model performance and reducing bias caused by imbalanced data. In the context of maritime accident severity prediction, employing a diversified balancing strategy to enhance model accuracy and stability remains unexplored. One novelty of this study is to investigate how the data balancing techniques could improve the state of the art in the field of maritime accident severity prediction. It will make new contributions by not only improved predictive capabilities but also provision of robust technical support for accurate maritime accident forecasting, promoting efficient risk assessment and prevention

measures.

2.3. Applications of explainable AI in high-risk decision-making and risk analysis

The XAI refers to the capability of ML models to articulate their decision-making processes in a manner comprehensible to humans [35]. The urgency for XAI has grown due to the "black-box" nature of many ML models, which obscures their internal decision-making mechanisms.

In recent years, XAI has emerged as a crucial research topic, particularly in high-stakes, decision-sensitive fields such as traffic accident analysis and medical diagnosis [36]. Various XAI approaches have been developed to enhance the transparency and interpretability of ML models, including rule-based models, tree models, and BNs [7,37]. Rule-based models utilize "if-then" rule sets to construct decision logic, offering intuitive and straightforward interpretability [37]. Tree models employ a hierarchical structure to visualize decision-making processes, simplifying user comprehension [7]. BNs use probabilistic reasoning to illustrate dependencies between variables, providing insights into complex systems [12]. Additionally, model-agnostic interpretability methods, such as LIME and SHAP, have gained significant attention for their applicability across various ML models. LIME generates local explanations by building simplified surrogate models, while SHAP, based on cooperative game theory, quantifies each feature's contribution to the prediction results [35,36].

Yuan et al. [38] applied RF combined with SHAP to evaluate the safety of real-time traffic flow data in connected vehicles. Their findings highlighted the significant impact of lane differences on average speed in real-time safety assessments. Vimbi et al. [39] conducted a systematic review of XAI methods for Alzheimer's disease (AD) detection, focusing on the potential, advantages, and challenges of SHAP and LIME in this context. The study demonstrated that XAI plays a crucial role in enhancing the reliability of AI-based AD prediction. Zhang et al. [40] utilized SHAP to perform interpretability analysis on tree-based maritime accident prediction models, revealing that crew-related factors significantly influence shipping safety. Similarly, He et al. [41] identified key factors affecting ship detention using SHAP. Their results indicated that deficiencies in crew certification and watchkeeping practices are critical determinants of ship detention. Chen et al. [42] evaluated the reliability of ML predictions using XAI, explicitly operationalising SHAP and LIME, to provide traceable rationales for safety-critical decisions. Complementing this, Palar et al. [43] embed SHAP within regression-reliability workflows to surface contribution profiles, providing a template for using SHAP as part of quantitative engineering analysis. Foundational sensitivity-analysis work within the reliability literature has also formalised Shapley-effect measures for feature importance, strengthening the theoretical underpinnings of SHAP-style explanations in safety contexts [44].

XAI has been applied across various fields, underscoring its pivotal role in enhancing the transparency of decision-making and the interpretability of models. In the context of maritime accidents, characterized by their multifactorial, complex, and high-risk nature, traditional "black-box" models often fail to provide a clear rationale for decision-makers, undermining the reliability and validity of predictions. The integration of XAI technologies in this study contributes a new solution to addressing these limitations by effectively identifying and explaining key factors influencing accident severity. This empowers maritime managers with deeper insights into model predictions, facilitating more informed risk assessments and decision-making.

2.4. State of the art and contributions of this study

This study presents three contributions to the field of maritime accident prediction, offering new perspectives and deeper insights into maritime safety management. For each of the three contributions (i.e., N1-N3) outlined below, the state of the art will be first summarized and the research gap will be identified, followed by an explanation on how this study addresses the gap.

N1. Multidimensional approaches to maritime accident prediction: Multi-model-based high-dimensional data processing and accurate prediction

State of the art: Most previous studies on maritime accident prediction have primarily relied on a limited number of ML models, with some utilizing only a single model [45]. This constraint hampers the model's ability to handle complex multivariate and nonlinear relationships, leading to reduced prediction accuracy and stability. Furthermore, a single model may fail to capture the underlying patterns within the data, thereby compromising the reliability of the prediction results.

Solution: This study introduces eight well-established ML models, leveraging the strengths of multiple algorithms for comparative analysis and comprehensive selection. By conducting cross-validation across various models, this study significantly enhances prediction accuracy and robustness, effectively addressing the challenges of high-dimensional data, nonlinear relationships, and complex interactions in maritime accident prediction. The proposed method demonstrates notable advantages in improving model prediction accuracy and stability, providing more reliable technical support with superior generalization capability, and contributing to the advancement of the maritime accident prediction field.

N2. Data imbalance in maritime accident prediction: Model optimization based on advanced data balancing techniques

State of the art: Data imbalance has long been a significant challenge in maritime accident prediction, affecting model performance. Existing studies often overlook the impact of class imbalance and rely on traditional ML models for accident prediction [12]. As a result, models become susceptible to the dominance of the majority class, leading to a marked decline in the prediction accuracy for minority classes. This bias diminishes the model's effectiveness in identifying high-risk accidents, ultimately compromising the reliability of the prediction results.

Solution: This study introduces six advanced data balancing techniques to address the issue of data imbalance in maritime accident data. These techniques enable the model to better manage the challenges posed by class imbalance, improving the prediction accuracy for serious accidents. Specifically, by synthesizing data from underrepresented categories and adjusting class weights, this study significantly enhances the accuracy and stability of the model in predicting critical accidents. Furthermore, the application of these balancing strategies effectively mitigates model bias, ensuring more reliable prediction results and providing a robust technical foundation for accurate maritime accident prediction and risk warning.

N3. An explanatory framework for maritime accident prediction: A combined SHAP and LIME-based application

State of the art: While ML methods have demonstrated significant advantages in handling complex maritime accident data, their 'black-box' nature has notably limited the interpretability and transparency of the models in practical applications. In maritime accident prediction, decision-makers require a clear understanding of model predictions to make informed and rational decisions. However, many existing studies have inadequately addressed the need for model interpretability, leading to insufficient transparency in prediction results. This lack of transparency, in turn, undermines the effectiveness of decision support [11]. Furthermore, traditional interpretation methods typically focus solely on global feature contribution analysis, overlooking the detailed examination of individual prediction results. This approach limits the comprehensive understanding of the decision-making mechanisms within the model.

Solution: This study develops an innovative interpretability framework that combines the SHAP and LIME approaches. By conducting global feature contribution analysis using SHAP, this study highlights the importance of individual features in the overall model prediction, enhancing the overall understanding of the model's decision-making process. Simultaneously, LIME provides local interpretations for

individual samples, enabling a deeper analysis of the drivers behind specific predictions and offering a clear basis for decision-making. The framework not only improves model interpretability but also enhances its transparency in practical applications, offering a novel perspective on decision support for maritime accident prediction models.

3. Materials and methodology

This study addresses the prediction of maritime accident severity through a series of systematic methods and techniques. Firstly, data integrity and consistency are ensured by collecting and preprocessing maritime accident data. Secondly, to address category imbalance in the dataset, various data balancing techniques are applied to enhance model performance across different categories. Thirdly, multiple ML models are then used for modelling, combined with methods such as hyperparameter optimization and cross-validation to further improve performance and prediction accuracy. Finally, to achieve model interpretability, explanatory AI techniques such as SHAP and LIME are introduced to provide an in-depth analysis of the impact of features on prediction results. The methodology framework is shown in Fig. 1. This section provides a detailed description of the materials and methods used to support the construction and validation of the maritime accident prediction model.

3.1. Materials

In this study, maritime accident investigation reports spanning from 2000 to 2019 were sourced from the repositories of seven internationally recognized maritime authorities. The detailed list of agencies is presented in Fig. 2. These seven data sources are the most representative ones supporting maritime accident studies in the literature, evident by the core journal papers relating to "maritime accidents" on Web of Science [2]. These data sources present diverse perspectives on accident causality and contributing to a robust dataset for analysis. Fig. 2 illustrates the sources and distribution of these maritime accident investigation reports. The distribution of reports across agencies indicates varying levels of contribution, with the ATSB accounting for the largest share at 26%, followed by the MAIB at 20%, and the BSU and NTSB at 13% and 12%, respectively. The China MSA and TSB each contribute 11%, while the JTSB provides 7%, forming a broad international dataset essential for this study's comprehensive analysis.

A thorough examination of the maritime accident reports contained within the specified database demonstrated substantial discrepancies in how information was presented in reports originating from various nations and regions. Some records were incomplete or biased, lacking essential information. To ensure data authenticity and completeness, this study implemented a rigorous screening process to remove records with missing key elements. Specifically, records that lacked descriptions of environmental conditions or the specific causes and consequences of accidents were excluded. Furthermore, to mitigate the impact of duplicate data on analysis accuracy, a strategy of cross-checking information was adopted, prioritizing the retention of more complete records. Through a rigorous data refinement and validation procedure, a well-structured collection of 1,294 systematically curated maritime accident investigation reports was obtained, providing robust support for this study. For details on the screening criteria and process, interested

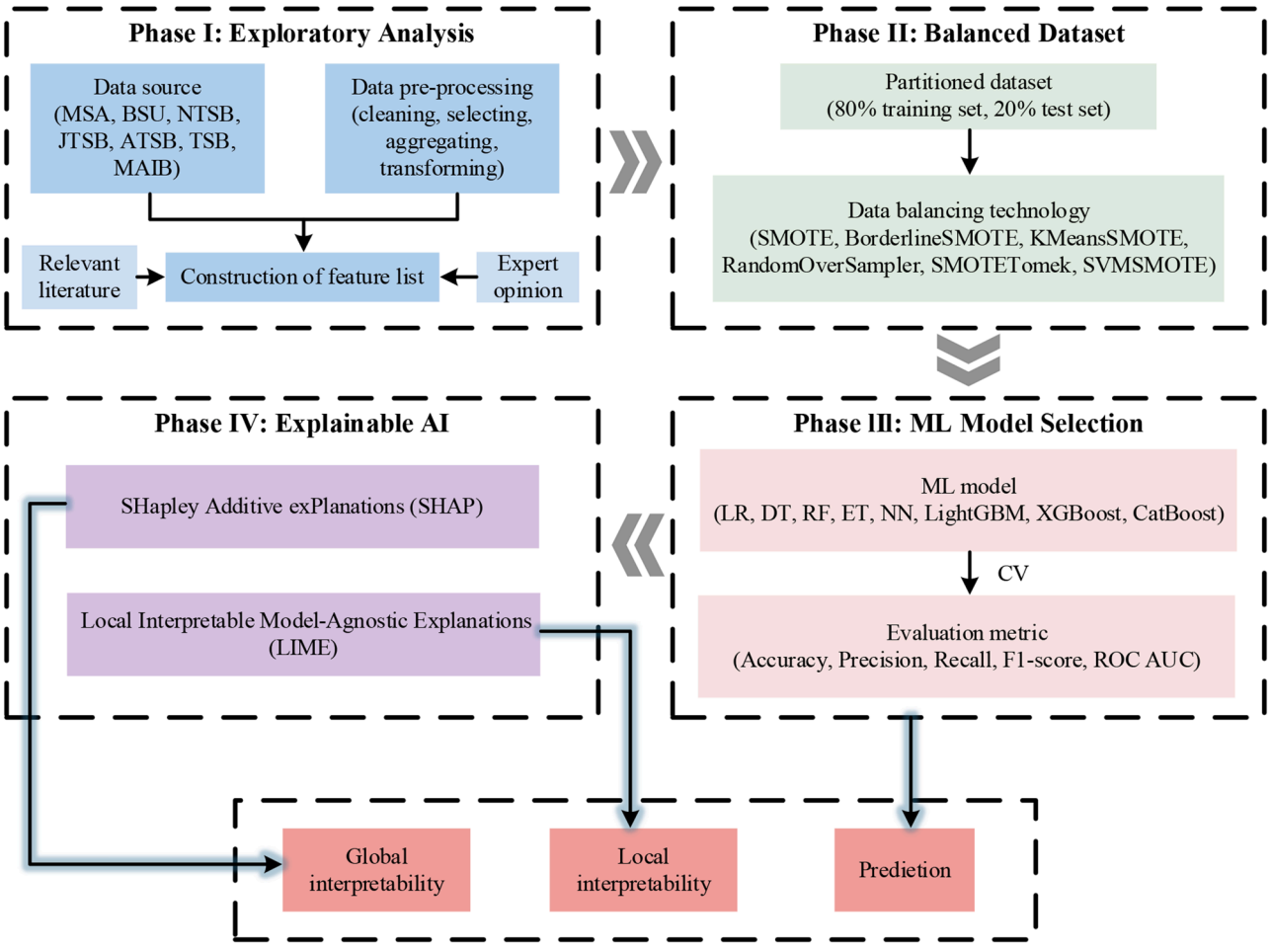


Fig. 1. The proposed prediction framework for maritime accident severity.

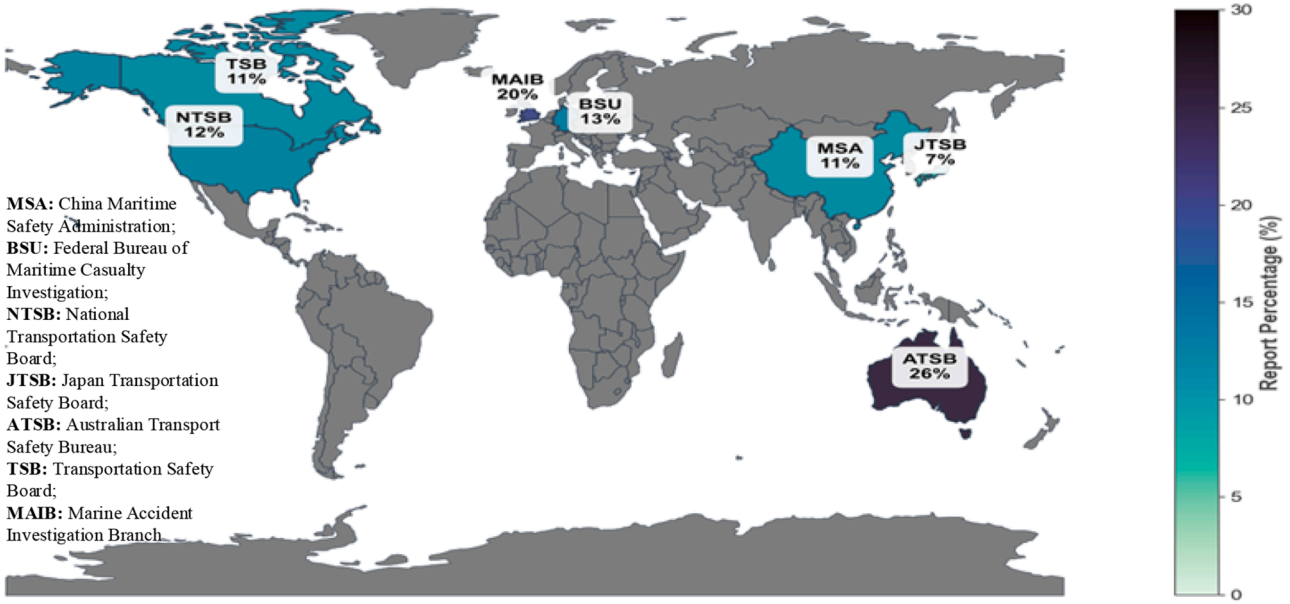


Fig. 2. Distribution of sources of marine accident investigation reports.

readers may refer to previous studies [1,2,12,46]. The construction of this dataset lays a solid foundation for subsequent analyses and ensures the validity and reliability of the findings.

3.2. Construction of feature list and dataset

In this study, a list of features encompassing multidimensional factors—such as accidents, human, ships, environment, and management—was developed from the perspective of system safety engineering based on relevant literatures [1,2,12,46], as shown in Table A of Appendix A. This list is intended to systematically analyse the key factors influencing the occurrence and severity of maritime accidents. To ensure a comprehensive analysis grounded in established maritime safety principles, this study retains the full set of curated features without performing a priori feature selection or dimensionality reduction. This approach is predicated on the understanding that each variable represents an empirically supported risk factor, and their premature removal could lead to the loss of critical explanatory information, thereby conflicting with the study's objective of conducting a holistic interpretability analysis.

In the field of maritime safety, the International Maritime Organization (IMO) has refined its classification of maritime accidents by assessing the extent of damage to ships, casualties, and ecological contamination. Specifically, the IMO categorizes the severity of maritime accident into three levels: less severe, severe, and very severe. Accidents involving casualties or significant ecological contamination (e.g., oil spills exceeding 500 tonnes) are classified as very serious accidents [46]. The data used in this study were derived from maritime accident investigation reports across multiple regions. In the process, this study encountered challenges in data standardization due to inconsistencies in reporting formats, classification standards, and terminology across regions, particularly in ensuring data consistency and enabling cross-regional comparisons. To address these issues and ensure the consistency of analytical results, as well as to prioritize reducing the risk of human casualties and ecological contamination in prospective maritime operations, this study adopts a streamlined scheme to classify the severity of maritime accidents. Specifically, accidents are categorized based on the most consistently reported indicators of severity—casualties and ecological contamination. Accidents involving either casualties or pollution are classified as serious, while those without are deemed non-serious. This study acknowledges that significant economic losses can also indicate severe accidents; however, due to inconsistencies in economic loss data across the seven sources and the IMO's three-level framework, the classification prioritizes casualties and pollution to maintain comparability. This streamlined classification emphasizes the critical need to prevent incidents resulting in human casualties or ecological contamination, while also highlighting the potential value of predicting casualty levels and pollution severity for enhanced risk management. Ultimately, the dataset contains 960 serious maritime accidents, accounting for 74.19%, and 334 non-serious maritime accidents, representing 25.81%. By analysing these variables, the objective of this study is to develop an effective classification and prediction model to identify and assess accident severity, providing a scientific basis for relevant stakeholders and promoting the ongoing improvement of maritime safety management and risk prevention.

3.3. Data balancing technologies

In data-driven modelling of maritime accident severity prediction, data balancing is a crucial prerequisite for ensuring model predictive performance [2,47]. Due to the class imbalance in the maritime accident dataset, where the percentage of serious accidents significantly exceeds that of non-serious accidents, the direct application of traditional ML models may lead to classification bias, particularly performance degradation in predicting the less frequent category (non-serious accidents) [25]. To address this, this study utilizes various data balancing techniques during data preprocessing to correct the bias in category distribution. By employing appropriate oversampling and undersampling methods, this section aims to enhance the model's ability to identify accidents in the underrepresented category while maintaining data representativeness, thus providing a foundation for the accurate prediction of accident severity.

Initially, a comprehensive review of all data balancing techniques reported in the literature was conducted. The selection process was grounded in three core criteria: (1) demonstrated empirical effectiveness in mitigating class imbalance within accident prediction tasks or analogous domains; (2) theoretical compatibility with the intrinsic characteristics of maritime safety data, particularly its skewed distribution and heterogeneous feature boundaries; and (3) computational feasibility and scalability for processing large, multidimensional datasets. Conversely, techniques lacking empirical support or deemed excessively complex were excluded. Consequently, six methods were chosen, comprising a

balanced assortment of oversampling, undersampling, and hybrid approaches. These include methods capable of capturing local decision boundary complexities (e.g., BorderlineSMOTE, SVMSMOTE) [48,49], those enhancing global data representativeness through interpolation or clustering mechanisms (e.g., SMOTE, KMeansSMOTE) [49,50], and techniques integrating noise reduction with boundary clarification (e.g., SMOTETomek) [51,52]. The inclusion of RandomOverSampler is justified by its methodological simplicity, low computational cost, and frequent application in class imbalance studies, making it a representative and practical choice among commonly used oversampling techniques [49]. To facilitate comparison and practical application, Table 1 presents the core principles of each technique, its applicable scenarios, and its advantages and disadvantages in the classification task, in a tabular format. This structured comparison provides a clear guideline for selecting appropriate data balancing methods and ensures more robust and fair classification performance when predicting maritime accident severity. The effective implementation of the aforementioned data balancing techniques not only mitigates the bias arising from class imbalance but also provides a robust basis for subsequent machine learning model training.

Before implementing data balancing, the dataset was first split into 80% for training and 20% for testing. This approach not only effectively mitigates the risk of data leakage but also ensures that the test set retains the original category distribution, reflecting the model's performance in real-world applications. The rationale behind this division will be discussed in detail in Section 3.4.2. With this process design, this study establishes a solid foundation for subsequent model training and evaluation, enhancing the reliability and validity of the classification task.

3.4. Machine learning modelling

After applying data balancing techniques effectively, the dataset is now optimally structured for predictive modelling. This section outlines the implementation of selected machine learning algorithms for maritime accident severity prediction. The methodology includes rigorous parameter tuning, advanced strategies to improve model generalization, and comprehensive performance evaluation metrics to ensure robust and reliable predictions in real-world scenarios.

3.4.1. Models selection

In predictive modelling of maritime accident severity, selecting an appropriate advanced ML model is crucial. In order to conduct a thorough assessment of various models' performance with the balanced dataset, eight widely recognized ML algorithms were employed, as listed in Table 2, are presented in this section. Each model has distinct characteristics and offers unique perspectives for predicting maritime accident severity. The eight ML models in this study were selected based on an extensive review of maritime and other transport accident prediction research. The inclusion criteria required that a model 1) demonstrated strong predictive performance in previous maritime or transport accident studies and 2) was implemented in widely used ML libraries to ensure reproducibility. Models that failed to meet these criteria or were computationally prohibitive were excluded.

To further improve the prediction performance of the models, this study employs a grid search technique for hyperparameter optimization of all selected ML models. By systematically exploring the hyperparameter space, grid search helps identify the optimal parameter combinations, thereby enhancing model performance during both training and validation phases. Coupled with hyperparameter optimization, this approach strengthens the model's learning ability, leading to more reliable predictions. This optimization process ensures that this study fully leverages the potential strengths of each model in constructing maritime accident severity prediction models, resulting in improved performance and prediction accuracy. A schematic diagram of

Table 1
The six data balancing techniques used in this study.

Technology	Theory	Applicable Scenario	Advantage	Disadvantage	Reference
SMOTE	The sample size of the minority class is increased by generating additional samples between existing minority samples using an interpolation algorithm.	This approach is suitable when there is a significant imbalance between majority and minority class distributions, yet the data exhibit homogeneity in their characteristics.	This method effectively increases minority class samples, enhancing the model's ability to recognize the minority class.	Insufficient control over sample generation at data boundaries can lead to misclassification.	[13,49, 50,]
BorderlineSMOTE	New samples should primarily be generated along the decision boundary, with an emphasis on the minority class samples that are likely to cause confusion.	This approach is suitable for minority class sample distributions characterized by ambiguous boundaries.	Enhance classification effectiveness at decision boundaries while minimizing the risk of misclassification.	If the boundaries are complex or noisy, the introduction of anomalous samples may occur.	[49,53]
KMeansSMOTE	Clustering a limited number of sample classes generates new samples within each cluster, thereby enhancing sample diversity.	This approach is suitable for a limited number of categories where the sample distribution exhibits greater heterogeneity.	The capacity to capture the intrinsic structure of a limited number of class samples facilitates the generation of new, more representative samples.	The effectiveness of clustering is contingent upon the data distribution and the selection of the number of clusters, with a corresponding high computational complexity.	[49,54]
RandomOverSampler	Randomly replicating samples from a limited number of classes increases their quantity, thereby creating a balanced distribution.	This approach is suitable for scenarios in which the data structure is simple and a limited number of sample classes can be easily replicated.	The method is straightforward, incurs low computational overhead, and is applicable across a wide range of scenarios.	This can easily result in overfitting problems, particularly when the minority class samples are limited.	[49,55]
SMOTETomek	Combining SMOTE-generated samples with the Tomek Link method allows for the removal of noisy data points and enhances the clarity of data boundaries.	This approach is suitable for situations where the data boundaries are ambiguous and there is a small amount of noise.	Additionally, this approach increases the number of class samples and reduces data noise, thereby enhancing classification effectiveness.	The removal of Tomek Link data may lead to the loss of some valid samples, thereby affecting the representation of the minority class.	[51,52, 56]
SVMSMOTE	Support Vector Machine (SVM)-based boundaries generate new samples to optimize the classification decision surface.	This approach is applicable to problems where the decision boundary is well-defined, but there are a limited number of categories with deficiencies.	This method enhances the SVM decision boundary, particularly by improving classification accuracy at the boundary.	This approach exhibits higher computational complexity, is sensitive to parameter selection, and requires increased computational resources.	[48,57]

Table 2
Eight well-established machine learning models used in this study.

Model	Core principle	Advantage	Applicable Scenario	Performance characteristics	Reference
LR	A linear model utilizing logistic functions designed for binary classification problems.	It is simple to implement, easy to interpret, and well-suited for small datasets.	Appropriate for datasets exhibiting a linear relationship between variables.	Performs well for linearly differentiable problems, but is sensitive to nonlinear data.	[58]
DT	Decisions are made through a tree structure that splits the data based on feature conditions.	The visualization is intuitive and accommodates non-linear relationships and missing values.	Suitable for small to medium-sized datasets with complex relationships among features.	Prone to overfitting, particularly at greater depths.	[59]
RF	An ensemble learning method comprising multiple DT, classified through a voting mechanism.	It exhibits high accuracy and robustness while being resistant to overfitting.	Applicable to large, high-dimensional datasets in scenarios where features interact with one another.	It can effectively address category imbalance; however, it incurs high model complexity.	[60]
ET	A variant of RF that constructs decision trees by randomly selecting both features and samples.	It is computationally efficient, facilitates rapid training, and is suitable for large-scale datasets.	Applicable to scenarios that require rapid training and prediction.	Its performance is comparable to that of RF in certain cases; however, the training time is shorter.	[61]
NN	Modelling the structure of a biological neural network facilitates feature learning through multiple layers of nodes.	Possesses a powerful learning capability to capture complex non-linear relationships.	Well-suited for large-scale and complex datasets, including image and text data.	It requires extended training times and large amounts of data, and it is susceptible to hyperparameter selection.	[62]
LightGBM	An efficient implementation of Gradient Boosting Decision Trees (GBDT) enables the processing of large-scale datasets with improved performance and scalability.	Efficient memory utilization and rapid training for high-dimensional data.	Well-suited for application scenarios that demand efficient computation and rapid response.	Demonstrates strong performance on large-scale data and effectively addresses class imbalance.	[63]
XGBoost	An enhanced GBDT that optimizes the model's training process.	Efficient and accurate, it supports parallel computation and feature selection.	It is widely utilized in various data science competitions and complex classification tasks.	It effectively handles missing values; however, parameter settings can be complex.	[64]
CatBoost	A gradient boosting algorithm optimized specifically for categorical features.	The user-friendly treatment of categorical variables and the automatic management of missing values reduce the risk of overfitting.	This approach is suitable for datasets with a large number of categorical features, particularly those exhibiting class imbalance.	The training process is fast and typically achieves strong performance across various datasets.	[65]

Note: LR is logistic regression; DT is decision tree; RF is random forest; ET is extra trees; NN is neural network; LightGBM is light gradient boosting machine; XGBoost is extreme gradient boosting; CatBoost is categorical boosting.

the grid search is provided in Fig. 3, visually demonstrating how the process works and contributes to performance enhancement. In this process, a discrete grid of possible values is first defined for each hyperparameter. The grid search method then systematically evaluates all possible combinations of these values by training and validating the model multiple times. Performance metrics are used to assess each configuration, and the combination yielding the best results is selected for final model deployment.

3.4.2. Split dataset

In the process of ML model training and evaluation, the division of the dataset is crucial. A reasonable ratio of training to testing data enhances the credibility of model evaluation and makes the prediction results more practical and informative. In data science, the 80%/20% division strategy is widely applied to various ML tasks, demonstrating its effectiveness in balancing model training and validation [66–68]. In this study, the maritime accident data was partitioned into 80% for training and 20% for testing. This division aligns with the commonly accepted industry practice in maritime accident severity research [2], providing sufficient training data for the model to capture potential patterns and relationships, while also ensuring adequate independent data to assess the model's generalization ability. This segmentation approach helps avoid scenarios where the model demonstrates strong performance on the training set but fails to generalize effectively to unseen data.

3.4.3. Cross-validation

Cross-validation is a powerful model evaluation technique for assessing the generalization ability of ML models. The basic principle is to evaluate model performance on unseen data by repeatedly dividing the dataset into subsets, thereby reducing the error introduced by data partitioning [2,7,10]. While traditional training/testing split methods may lead to instability in model performance evaluation, cross-validation provides more reliable results by ensuring that each subset serves as a validation set at some point.

In ML research, 5-fold or 10-fold cross-validation is commonly used to effectively control overfitting. Studies have shown that 5-fold cross-validation provides an optimal trade-off between computational cost and accurate assessment of model performance, while 10-fold cross-validation requires more computational resources and time but does not offer a significant performance advantage over 5-fold cross-validation [69,70]. Therefore, based on the sample size and findings from related literature [2,7], this study opted for 5-fold cross-validation.

Specifically, in 5-fold cross-validation, the entire dataset is randomly divided into five equal-sized subsets. In each fold, one subset is selected as the validation set, and the remaining four subsets are used for model training. This process is repeated five times, each time using a different subset as the validation set, ensuring that each subset is used for validation once and that the model is trained on the entire dataset. This strategy ensures that the model is evaluated on each subset, providing a more comprehensive test of its generalization ability. By applying 5-fold cross-validation, five folds, each with five sub-datasets, are generated to test and analyse model performance, as shown in Fig. 4. The average score across these five folds is used as an indicator of the model's final performance, significantly reducing overfitting and sample bias during the learning process. This cross-validation technique ensures that the results of the maritime accident severity prediction model are highly reliable, providing a strong scientific basis for decision-makers.

3.4.4. Evaluation metrics

In the evaluation of maritime accident severity prediction models, selecting appropriate evaluation metrics is essential. This study employs several key metrics—Accuracy, Precision, Recall, F1-score, and ROC AUC—which provide a multi-dimensional view of model performance [71]. To obtain a thorough insight into the model's performance across different categories, the confusion matrix is introduced as a vital tool for visualizing the relationship between model predictions and actual

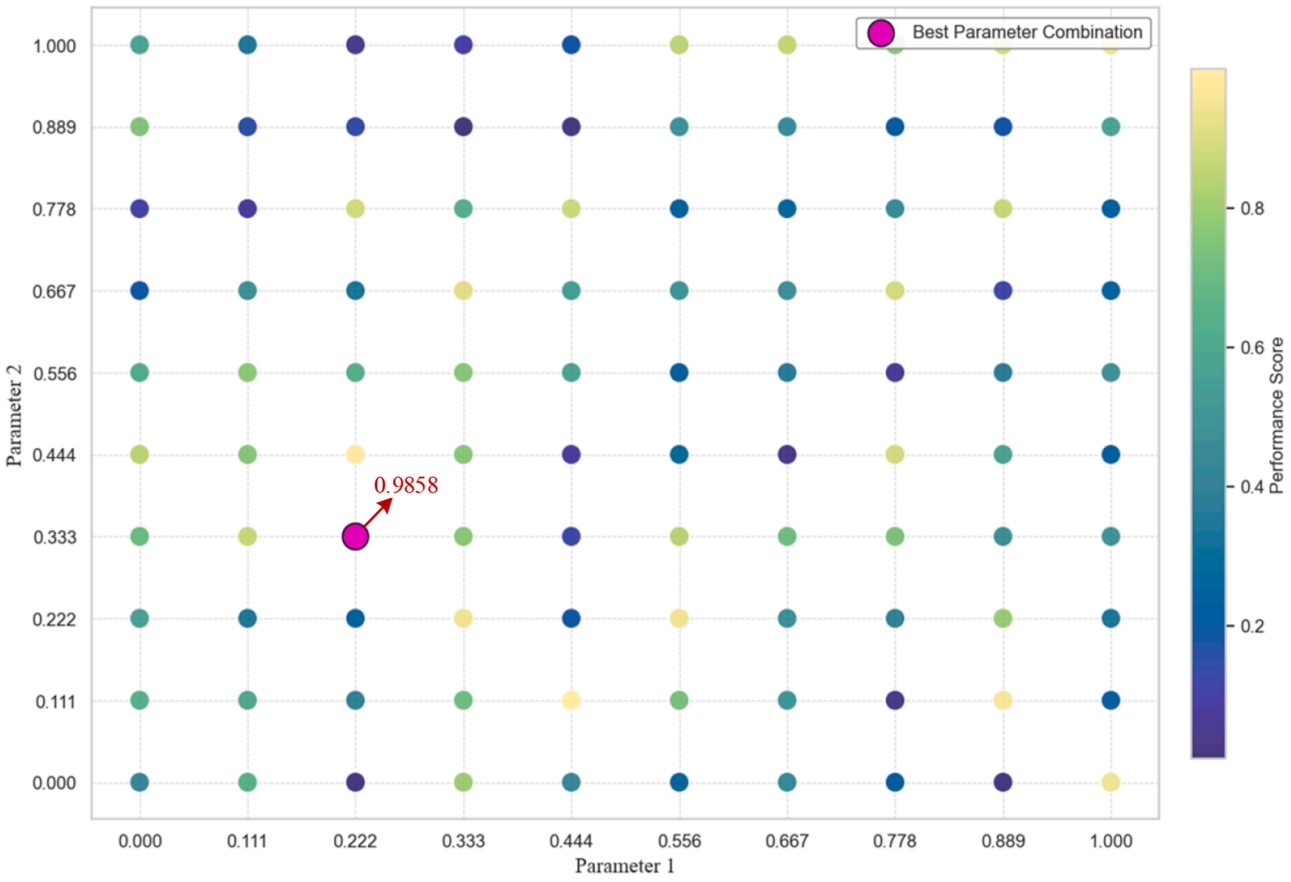


Fig. 3. The grid search for hyperparameter optimization.

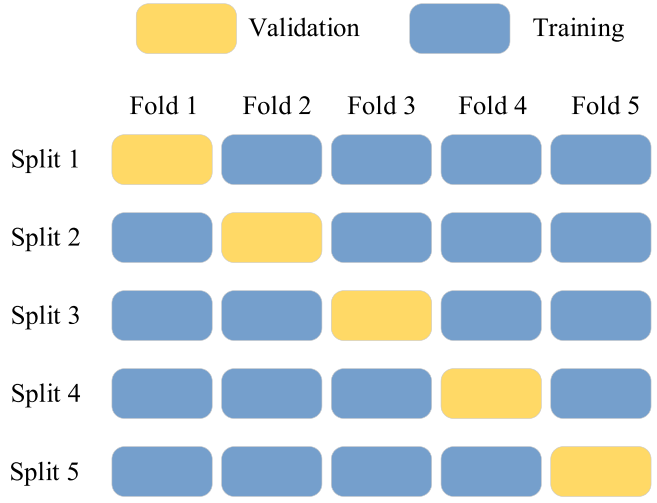


Fig. 4. The 5-Fold cross-validation process.

outcomes. As shown in Fig. 5, the confusion matrix offers an in-depth analysis of classification performance through four key components: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Table 3 provides a detailed explanation of these evaluation metrics, the rationale behind their selection, and further emphasizes their significances in evaluating model performance, particularly in the context of the confusion matrix.

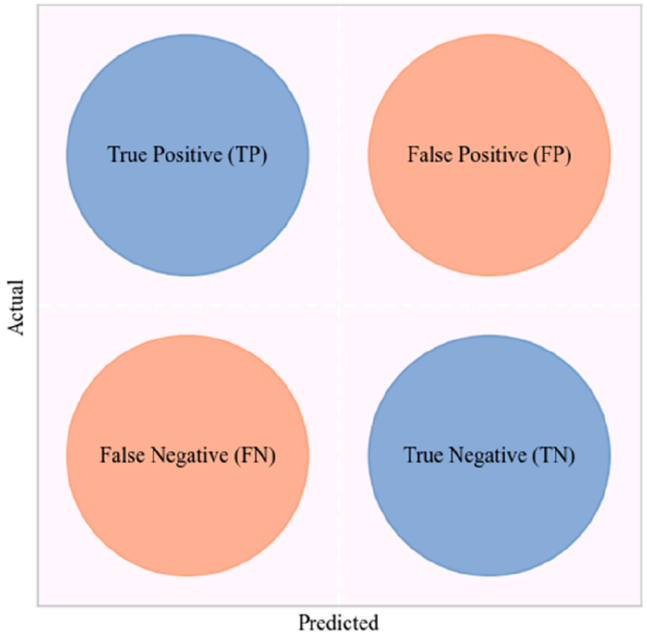


Fig. 5. The confusion matrix.

3.5. *Interpretability of results generated by ML*

In the application of ML models, the interpretability of results is crucial to enhance both the model's credibility and the effectiveness of its application [36,72]. To acquire a comprehensive understanding of

Table 3
Summary of the metrics.

Metric	Definition	Reason for choosing the metric
Accuracy	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	Assessing the accuracy of the overall predictions ensures that the model's ability to classify all categories of maritime accident severity is relatively balanced.
Precision	$\text{Precision} = \frac{TP}{TP + FP}$	To accurately reflect the models' effectiveness in predicting serious accidents, it is essential to reduce the false positive rate, ensuring efficient allocation of resources.
Recall	$\text{Recall} = \frac{TP}{TP + FN}$	Evaluate the model's ability to identify actual serious accidents and ensure the timely detection of potentially risky events in maritime safety management.
F1-score	$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	Considering precision and recall rates together is crucial for providing an assessment that balances accuracy and coverage in the recognition of serious accidents.
ROC AUC	$\text{ROC AUC} = \int_0^1 \text{TruePositiveRated}(\text{FalsePositiveRate})$	Evaluating the model's ability to classify across different thresholds reflects its overall capacity to discriminate in predicting maritime accident severity.

the decision-making mechanism behind the maritime accident severity prediction model, this study employs the SHAP method for global interpretation and the LIME method for local interpretation for individual samples. Fig. 6 illustrates the decision boundary of a nonlinear black-box model and its outputs for a specific maritime accident instance. The pink and green areas in Fig. 6 represent the model's different predicted outcomes for accident severity, with the decision boundary clearly distinguishing between non-serious and serious accidents. This visual representation not only intuitively demonstrates the model's decision-making process but also provides important insights into how the model classifies accidents within various feature spaces. Specific cases are highlighted, showing their positions within the feature space and the model's corresponding predictions. By applying SHAP, this study assesses the global feature importance of the model, revealing key factors that influence accident severity. Simultaneously, LIME offers local explanations for individual cases, allowing a deeper exploration of the model's decision-making rationale in specific instances. This dual interpretation framework significantly enhances model transparency and aids decision-makers in making more accurate and reliable judgments in practical applications. Ultimately, it improves understanding of the model and provides a solid theoretical foundation for subsequent risk management and decision support.

3.5.1. *Shapley additive explanations (SHAP)*

In maritime accident severity prediction, SHAP values are commonly used to quantify the influence of features on model output. Based on cooperative game theory, this method determines the contribution of each feature to a particular prediction by computing its average marginal influence across all possible feature combinations [73]. In this study, SHAP values are employed to interpret the contribution of each individual feature in maritime accident severity prediction models. After training the models, SHAP analysis quantifies the influence of each feature on the model's prediction outcomes. A positive SHAP value indicates a higher likelihood of predicting a severe accident, while a negative SHAP value suggests a mitigating effect. The explanatory power makes SHAP a crucial tool for complex ML models, particularly in maritime safety, where a deeper understanding of decision-making processes is essential. This approach helps identify critical risk factors associated with maritime accidents.

By analysing the SHAP values, this study not only provides insights into the relative importance of each feature in the model's decision-

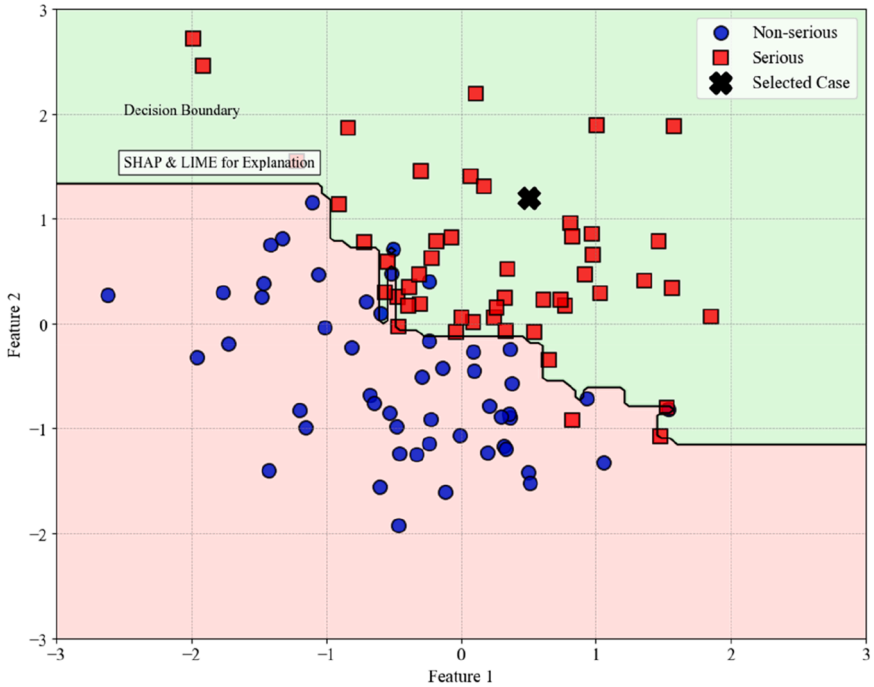


Fig. 6. Decision boundary of a non-linear ML model for maritime accident severity prediction.

making, but also offers practical recommendations for maritime safety management. For example, by identifying the key factors influencing the prediction of serious accidents, decision-makers can design more targeted interventions to improve overall safety. The process of calculating SHAP values can be expressed by Eq. (1).

$$x_j = \sum_{R \subseteq M\{j\}} \frac{|R|!(|M| - |R| - 1)!}{|M|!} [f(R \cup \{j\}) - f(R)] \quad (1)$$

where x_j is the SHAP value of the j^{th} feature.M is the complete set of features, representing all features in the model, and R is a subset of M excluding feature j. R is used to evaluate the marginal contribution of feature j to the model's output under various feature combinations. $f(.)$ is the prediction function of the model. $f(R)$ represents the predicted result when only the features in subset R are used. $f(R \cup \{j\})$ represents the prediction result when feature j is added to R. Expression $f(R \cup \{j\}) - f(R)$ calculates the incremental contribution of feature j to the model's output for a given combination of features R, reflecting its marginal impact. $\frac{|R|!(|M|-|R|-1)!}{|M|!}$ is the weighting factor that represents the weights of all permutations containing feature j, ensuring that the contribution of feature$j$ is fairly calculated across all possible feature combinations. $|R|!$ and $(|M| - |R| - 1)!$ are the factorials of subset R and the remaining features, respectively, and $|M|!$ is the number of full permutations of the complete set of features.

3.5.2. Local interpretable model-agnostic explanations (LIME)

LIME is a tool designed to improve the interpretability of a model by enabling the interpretation of predictions from any classifier or regressor in an intuitive and easily understandable manner [36]. In this study, LIME is used to interpret the predictions of maritime accident severity models. After training the models, LIME constructs local surrogate models for individual predictions, identifying the most influential factors contributing to each outcome. This approach enables a case-specific analysis of maritime accidents, offering deeper insights into the model's decision-making process, improving the interpretability of predictive results, and identifying key factors influencing accident severity.

Unlike the global interpretability provided by SHAP, LIME focuses on local interpretability by examining how the model behaves in the vicinity of a specific instance. By constructing a local, interpretable linear model for each prediction, LIME identifies the features that most influence a given prediction, thereby clarifying the decision-making process. Specifically, LIME fits a simple interpretable surrogate model on the perturbed instances around the instance of interest using Eq. (2), which is to minimize the difference between the local surrogate model and the original model.

$$Y(y) = \text{argmin}_{k \in K} P(g, k, \pi_y) + \Omega(k) \quad (2)$$

where y denotes the instance to be interpreted, while g is the original complex, nonlinear predictive model. $Y(y)$ represents the process of finding an interpretable local surrogate model in the vicinity of a specific instance y. The interpretable model k serves as a simplified approximation of g, providing local interpretation in the neighbourhood of y to enhance comprehensibility. π_y refers to the proximity metric between instance y and the data instances used to learn the explanatory model. In this study, Euclidean distance is used as the metric. $P(g, k, \pi_y)$ measures the fidelity of the interpretable model k to the original model g in the neighbourhood π_y. Typically, this fidelity is assessed by the Euclidean distance between the predictions of g and k within the neighbourhood π_y. The definition of the neighbourhood π_y is crucial, as it ensures that k approximates the behaviour of g only within a local region around y, without requiring an exact fit over the entire input space. The complexity term $\Omega(k)$ introduces a simplification constraint on model k, encouraging it to remain as simple as possible. By minimising $\Omega(k)$, LIME aims to generate models that are both accurate and interpretable within the local region π_y. Thus, $Y(y)$ seeks an optimal interpretable model that balances fidelity (measured by $P(g,k,\pi_y)$) and interpretability (controlled by $\Omega(k)$). Once model k is obtained, its coefficients or feature importance values are extracted to explain which features contribute most to the original model g for an individual instance.

4. Findings and discussion

4.1. Construction and evaluation of a maritime accident severity prediction model

In this section, six data balancing techniques are utilized, including SMOTE, BorderlineSMOTE, KMeansSMOTE, RandomOverSampler, SMOTETomek, and SVMSMOTE. These techniques are designed to handle imbalanced datasets and improve the representativeness of minority class samples. Additionally, eight advanced ML models—LR, DT, RF, ET, NN, LightGBM, XGBoost, and CatBoost—are used. By combining these six data balancing techniques with the eight ML models, a total of 48 model combinations are constructed, and a comprehensive performance evaluation is conducted for each of these combinations.

4.1.1. Data balancing techniques and model performance evaluation

In this sub-section, the impact of data balancing techniques on the models' predictive capabilities will be evaluated. The results of the performance metrics of the eight ML models without data balancing techniques are provided in Table B2 of Appendix B. These results provide a baseline for subsequent comparative analyses with various data balancing techniques.

Table 4 presents the performance of various ML models after incorporating multiple data balancing techniques, which is evaluated by accuracy, precision, recall, F1-score, and ROC AUC. As shown in Table 4, the performance of the ML models is significantly improved after applying multiple oversampling methods. For example, comparing with the results in Table B2 of Appendix B, with the application of SMOTE, the RF model's accuracy increases from 0.7699 to 0.8392, recall increases from 0.6822 to 0.8479, and ROC AUC increases from 0.8421 to 0.9180; and the CatBoost model's accuracy improves from 0.7759 to 0.8479 and ROC AUC increases from 0.8404 to 0.9209. These findings indicate that data balancing techniques significantly improve model performance under skewed data conditions, as reflected in enhanced prediction accuracy. This observation aligns with the theoretical foundations of related studies [25,47,74] and highlights the importance of effective data handling strategies in maritime accident severity prediction.

Moreover, we investigate how different data balancing methods impact model performance to identify the optimal model combinations. To facilitate visual comparison, a colour coding scheme is employed in Table 4: green indicates lower performance values, yellow represents medium performance, and orange-red identifies high performance values. This colour coding enhances the readability of Table 4 and enables the reader to quickly identify the model with the best performance under each data balancing approach. As shown in Table 4, the two data balancing methods, RandomOverSampler and SMOTETomek, exhibit orange-red areas across multiple metrics for several models, suggesting that they yield superior performance. The combination of RandomOverSampler and CatBoost demonstrates the highest predictive performance when considering all evaluation metrics. Specifically, the RandomOverSampler-CatBoost model achieves an accuracy of 0.8645, a precision of 0.8438, a recall of 0.8970, a F1-score of 0.8681, and a high ROC AUC value of 0.9369. Upon examining the colour coding in Table 4, it is evident that this combination is highlighted in orange-red across all evaluation metrics, emphasizing its significant advantage in handling unbalanced datasets.

The excellent performance of the RandomOverSampler-CatBoost combination can be attributed to several factors. On the one hand,

Table 4
The performance of machine learning models with data balancing.

Data balancing techniques	Model	Accuracy	Precision	Recall	F1-score	ROC AUC
SMOTE	LR	0.7797	0.7898	0.7654	0.7764	0.8500
	DT	0.7869	0.7982	0.7813	0.7857	0.8618
	RF	0.8392	0.8346	0.8479	0.8405	0.9180
	ET	0.8265	0.8225	0.8336	0.8312	0.9155
	NN	0.7798	0.7684	0.8178	0.7621	0.8663
	LightGBM	0.8289	0.8251	0.8369	0.8278	0.9116
	XGBoost	0.8344	0.8337	0.8368	0.8329	0.9140
	CatBoost	0.8479	0.8394	0.8622	0.8490	0.9209
BorderlineSMOTE	LR	0.7393	0.7463	0.7257	0.7356	0.8151
	DT	0.7710	0.7768	0.7638	0.7699	0.8306
	RF	0.8407	0.8289	0.8621	0.8462	0.9130
	ET	0.8249	0.8256	0.8558	0.8371	0.9144
	NN	0.7718	0.7806	0.8257	0.7428	0.8363
	LightGBM	0.8297	0.8169	0.8527	0.8327	0.9108
	XGBoost	0.8273	0.8209	0.8400	0.8294	0.9086
	CatBoost	0.8439	0.8288	0.8685	0.8471	0.9165
KMeansSMOTE	LR	0.8014	0.8201	0.7756	0.7898	0.8879
	DT	0.7943	0.8017	0.7774	0.7837	0.8725
	RF	0.8393	0.8612	0.8167	0.8294	0.9200
	ET	0.8251	0.8410	0.8120	0.8208	0.9183
	NN	0.7950	0.8184	0.8087	0.7855	0.8827
	LightGBM	0.8283	0.8353	0.8215	0.8227	0.9143
	XGBoost	0.8409	0.8595	0.8166	0.8319	0.9183
	CatBoost	0.8433	0.8520	0.8341	0.8380	0.9190
RandomOverSampler	LR	0.7773	0.7794	0.7750	0.7768	0.8418
	DT	0.7829	0.7904	0.7670	0.7771	0.8455
	RF	0.8621	0.8475	0.8860	0.8664	0.9344
	ET	0.8487	0.8493	0.8543	0.8463	0.9372
	NN	0.7853	0.7966	0.8320	0.7986	0.8592
	LightGBM	0.8455	0.8249	0.8781	0.8489	0.9222
	XGBoost	0.8400	0.8198	0.8717	0.8434	0.9242
	CatBoost	0.8645	0.8438	0.8970	0.8681	0.9369
SMOTETomek	LR	0.7880	0.7998	0.7707	0.7838	0.8599
	DT	0.8045	0.8088	0.8004	0.8028	0.8641
	RF	0.8532	0.8502	0.8615	0.8440	0.9283
	ET	0.8416	0.8449	0.8468	0.8408	0.9278
	NN	0.8020	0.8010	0.8153	0.7904	0.8776
	LightGBM	0.8523	0.8432	0.8681	0.8532	0.9235
	XGBoost	0.8507	0.8490	0.8532	0.8492	0.9238
	CatBoost	0.8606	0.8552	0.8714	0.8611	0.9335
SVMSMOTE	LR	0.7528	0.7591	0.7400	0.7481	0.8262
	DT	0.7734	0.8033	0.7257	0.7614	0.8439
	RF	0.8463	0.8366	0.8542	0.8401	0.9135
	ET	0.8297	0.8302	0.8463	0.8348	0.9154
	NN	0.7718	0.7718	0.7100	0.7799	0.8396
	LightGBM	0.8233	0.8163	0.8385	0.8249	0.9067
	XGBoost	0.8320	0.8254	0.8448	0.8338	0.9091
	CatBoost	0.8486	0.8345	0.8717	0.8516	0.9149

Note: All evaluation metrics (Accuracy, Precision, Recall, F1-score, and ROC AUC) are based on five-fold cross-validation. Green denotes a low value, yellow signifies a medium value, and orange-red represents a high value.

RandomOverSampler improves the model's ability to learn the minority class by randomly augmenting the number of minority class samples, thus enhancing the classifier's sensitivity to rare events. This technique is simple to implement and more computationally efficient compared to other complex data balancing methods such as SMOTE and KMeansSMOTE [75]. On the other hand, the CatBoost model offers significant advantages when handling categorical features, with its unique ordered boosting strategy effectively minimizing the likelihood of overfitting and enhancing the model's ability to generalize [76]. This characteristic is particularly beneficial in maritime accident data analysis, where the dataset often exhibits significant category imbalances and a high-dimensional feature space.

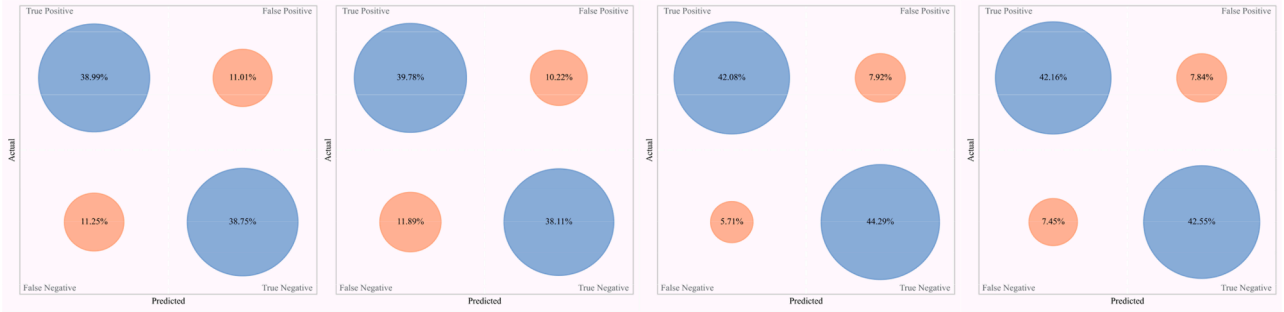
4.1.2. Confusion matrix analysis

The confusion matrix plays a crucial role in model evaluation, as it reflects not only the performance of the classification model but also reveals the prediction accuracy across different categories [7,10]. We calculate the confusion matrices for all the combinations of the eight ML models and six data balancing techniques. As an illustration, Fig. 7 shows the confusion matrix for eight ML models trained using the RandomOverSampler data balancing method. The confusion matrices of the other combinations are provided in Figs. C1-C5 of Appendix C. In Fig. 7, blue circles indicate accurate predictions, while yellow circles indicate incorrect predictions. The size of the circles is proportional to the number of predictions, with larger circles indicating a higher frequency of correct or incorrect predictions in that category. Consequently, larger blue circles signify superior model performance in that category. From Fig. 7, it is evident that the RandomOverSampler-CatBoost model demonstrates a significant advantage in the confusion matrix. The area of the blue circles for this model is substantially larger than those of the other models, illustrating its superior ability to correctly predict samples from smaller categories. This observation not only highlights the model's strong adaptability in complex data environments but also reinforces the importance of data balancing strategies in improving model performance.

In summary, the excellent performance of the RandomOverSampler-CatBoost model can be attributed to its effective data balancing strategy and the CatBoost model's strong adaptability in complex data environments. This study confirms the significant potential of data balancing techniques in enhancing the performance of maritime accident prediction models, particularly in identifying rare classes of samples. It demonstrates that classification accuracy can be significantly improved by appropriately selecting the data balancing method, providing more precise decision support for maritime safety management. The combination of RandomOverSampler and CatBoost increases the number of minority class samples (non-serious accidents) and enhances the model's sensitivity to rare events, while CatBoost's efficient algorithm excels in handling multi-dimensional feature spaces.

4.2. Combined SHAP and LIME analysis and key features identification

This study pioneers the combination of SHAP and LIME approaches to enhance the interpretability and reliability of key features influencing the severity of maritime accidents. Based on the optimal model identified in Section 4.1, the RandomOverSampler-CatBoost model, this subsection conducts an in-depth analysis of its internal decision logic and examines the specific impact of key features on predictions. SHAP and LIME each offer distinct strengths in model interpretation: SHAP excels at global feature contribution analysis, revealing the weights and relative importance of features in the overall model predictions, while LIME focuses on fine-grained, local interpretations of individual samples, providing deep insights into the decision logic for specific predictions [36,72]. This study effectively integrates both methods to maximize interpretative depth. Firstly, a global analysis using SHAP identifies the key features significantly impacting the model's predictions and clarifies their relative contributions. Then, local analyses using LIME further explore the model's decision logic at the sample level, offering precise explanations for specific predictions. Finally, to validate the reliability of local interpretations, this study compares the results from SHAP and LIME, enhancing the credibility of the analysis. This multi-level, integrated approach not only deepens the understanding of key features but also provides a solid foundation for developing targeted safety interventions, while improving both the prediction accuracy and interpretability of maritime risk assessment models.



(a) RandomOverSampler-LR (b) RandomOverSampler-DT (c) RandomOverSampler-RF (d) RandomOverSampler-ET

(e) RandomOverSampler-NN (f) RandomOverSampler-LightGBM (g) RandomOverSampler-XGBoost (h) RandomOverSampler-CatBoost

Fig. 7. The result of confusion matrix for various ML models with the RandomOverSampler data balancing technique.

4.2.1. Global analysis with SHAP

To comprehensively assess global feature importance in model prediction, this study quantifies each feature's contribution to model decision-making by aggregating its SHAP value. Specifically, by calculating the absolute values of these SHAP values and averaging them across the entire dataset, this study robustly reflects the overall impact of each feature on the model, regardless of whether its influence on the output is amplified or diminished [77]. This approach enables the objective identification of features with the greatest impact on predicted outcomes. Features are ranked based on the mean of these absolute SHAP values, with higher values indicating a more significant role in model predictions, providing a solid foundation for subsequent analysis and decision-making.

This study enhances the understanding of model prediction by demonstrating the application of SHAP values in feature contribution assessment through Figs. 8, 9, and 10. Fig. 8 provides a comprehensive SHAP-based visualization, illustrating the overall contribution of each feature to the model output. Fig. 9 highlights the nine most important features identified in the SHAP analysis, emphasizing their significance and influence on model predictions. Fig. 10 offers an in-depth assessment of how important features influence maritime accident modelling through interpretable SHAP dependency plots. The combination of these visualizations not only strengthens the quantitative evaluation of feature importance but also provides decision-makers with a clear visual guide to identify key features that impact model predictions.

Fig. 8 presents a combined SHAP global force diagram and SHAP heat map, analysing feature contributions to the model for predicting maritime accident severity at both the global and local levels. The top half of Fig. 8 displays the SHAP global force diagram, which illustrates each feature's overall contribution to model predictions across all samples. The magnitude and direction of each feature's influence on maritime accident severity predictions are indicated by colour (red for positive influence and blue for negative influence) and bar length. Larger SHAP values correspond to greater influence on the prediction. This global view reveals the most important features, their specific contributions, and interactions, providing in-depth insights into the model's decision logic and key drivers. The bottom half of Fig. 8 shows the SHAP heat map, which helps visualize the feature contributions for individual samples. Each sample's feature contribution is represented through a colour gradient (red for positive contribution and blue for negative contribution), highlighting the impact of features on individual predictions from a local perspective. Each column in the heat map represents a sample, and each row corresponds to a feature, with colour intensity reflecting the magnitude of the feature's contribution to that sample's prediction. This visualization enables the identification of model consistency or variability across samples, particularly the heterogeneity in the impact of key features on different samples. This dual visualization significantly enhances model interpretability by providing clear graphical representations of feature contributions at both the global and local levels. It not only aids in understanding the model's decision logic and feature importance, but also lays a solid foundation for further in-depth analysis and discussions, facilitating a comprehensive exploration of the specific impact of features on prediction outcomes.

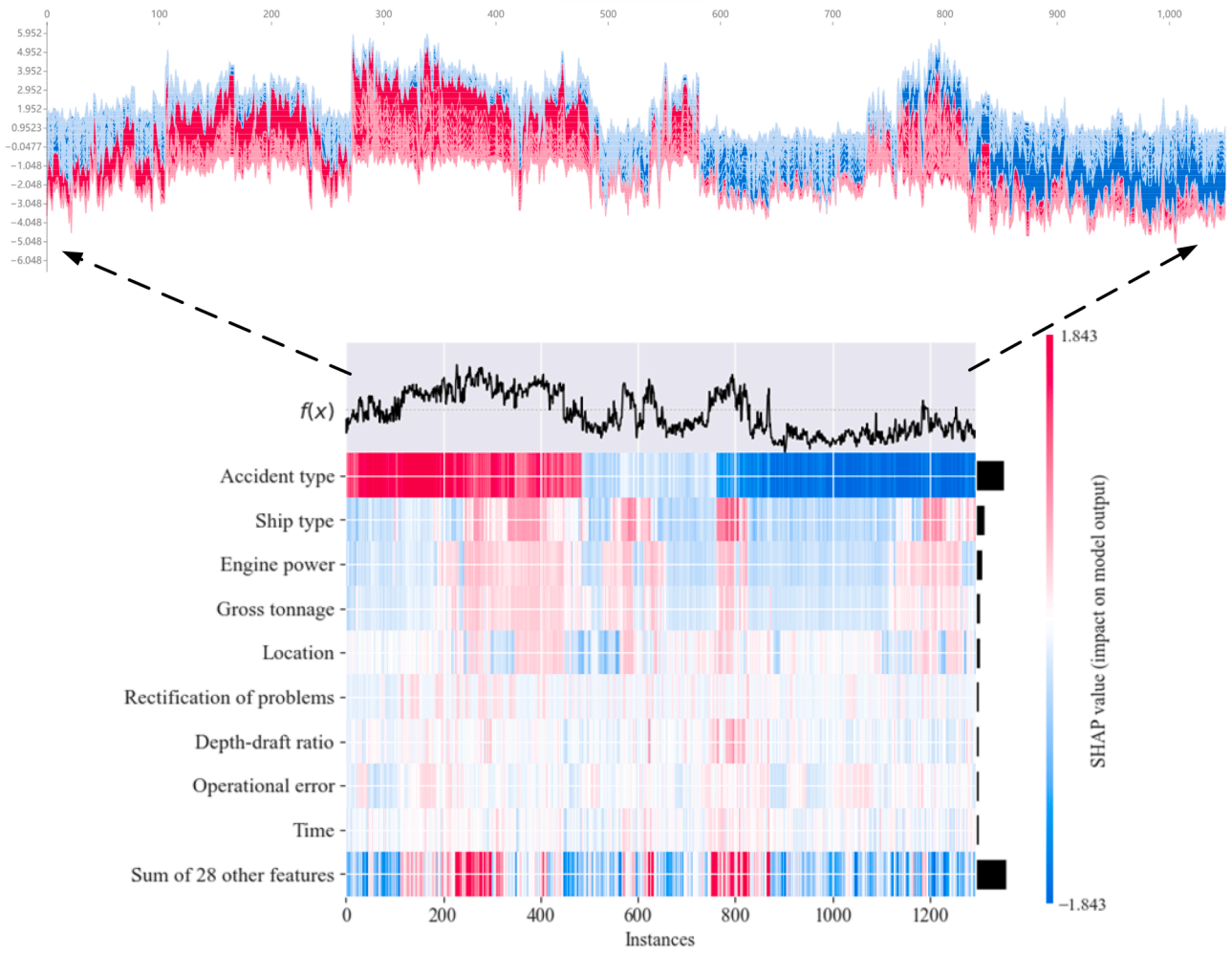
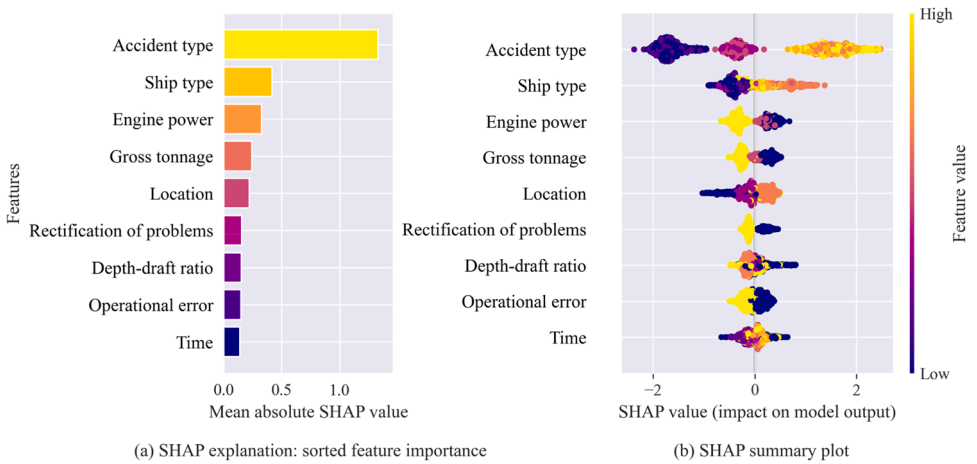


Fig. 8. Comprehensive SHAP-based visualisation of feature contributions.



(a) SHAP explanation: sorted feature importance (b) SHAP summary plot

Fig. 9. SHAP analysis of feature importance and impact on model predictions (Top 9 features).

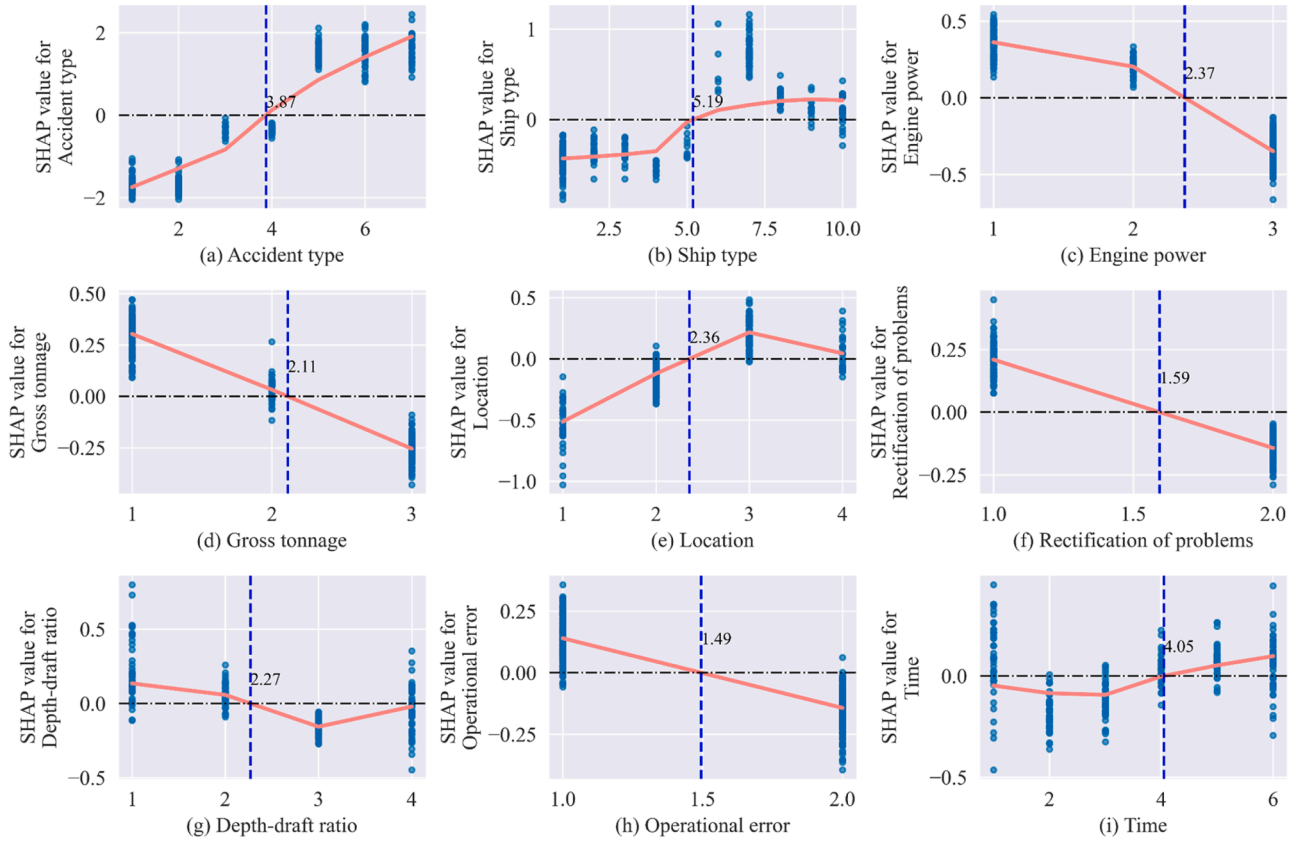


Fig. 10. Interpretable SHAP dependence plots for the top nine features.

To further investigate the importance ranking of each feature and its specific impact on prediction outcomes, this study presents the results of feature importance assessment based on SHAP analysis, as shown in Fig. 9. Fig. 9(a) highlights the nine features with the greatest impact on accident severity prediction, with accident type, ship type, engine power, and gross tonnage occupying the top four positions. These features make significant contributions to the predictive power of the model, consistent with findings in existing literature [10]. As a core attribute of the accident itself, accident type, which likely reflects the nature and potential consequences of the accident, is widely recognized as a key factor influencing accident severity [10,46]. The strong relationship between accident type and severity may be due to the unique characteristics of each accident type in terms of its occurrence mechanism, scope of impact, and emergency management, further supporting the importance of this feature in the model's predictions. Similarly, ship type, engine power, and gross tonnage, as key indicators of the ship's physical characteristics, highlight the potential influence of the vessel itself on accident severity. Significant differences exist between ship types in terms of operational modes, structure stability, and risk resistance. Engine power and gross tonnage, on the other hand, reflect the ship's power characteristics and load-carrying capacity, both of which can affect accident severity [7,10,12].

In Fig. 9(b), each point represents the SHAP value of a sample, illustrating the trend of feature values in predicting accident severity. The colour coding corresponds to the numerical magnitude of the feature, with yellow indicating a high feature value, purple a medium

value, and dark blue a low value. Positive and negative SHAP values represent the direction of the feature's contribution to accident severity. A positive value signifies a positive influence on the prediction of severity, while a negative value indicates a negative influence. By cross-referencing with Table A of Appendix A, a context-specific interpretation of the SHAP values can be obtained. Firstly, when the eigenvalues of maritime accident types are large, they typically correspond to high-severity accident types such as sinking and hull damage. This suggests that such accidents, characterized by their potentially severe consequences and high risks, significantly drive the prediction of high severity in the model. This trend aligns with the reality that sinking and hull-damage accidents often result in severe structural damage, casualties, and property loss, and thus carry significant weight in predictive models [12,78]. Furthermore, these high-risk accident types differ substantially from minor collisions or mechanical failures in terms of subsequent management and emergency response needs, which also enhances their influence on the prediction of accident severity. Secondly, when the eigenvalues for ship types are large, they generally correspond to ship types such as tugboats, cruise accident severity due to the specifics of their functions and operating environments. For example, tugboat operations are more prone to accidents due to the complexity of towing tasks; cruise ships present safety hazards due to the large number of passengers and activities on board; and fishing vessels face higher accident risks due to the unpredictability of the operating environment and suboptimal equipment [10,79]. Thirdly, as the ship's main engine power and gross tonnage decrease, their influence on accident severity intensifies. This is because ships with smaller engine power and tonnage tend to be more unstable in rough sea conditions, increasing the likelihood of serious accidents [2].

Fig. 10 provides a detailed illustration of the relationship between the influence of the nine core features and accident severity, including their threshold effects. Locally weighted scatterplot smoothing is applied to smooth the scatterplot and emphasize the marginal effects of the independent variables. The steeper the curve, the greater the marginal contribution of the features. The dashed line in Fig. 10 marks the value of a feature when the SHAP value is zero, representing the cut-off point where the feature has a "neutral" effect on the model output. These critical points serve as a scientific foundation for identifying the turning intervals and boundary effects of feature impacts, offering insights into the risk contribution thresholds of different features. In Fig. 10, the sensitivity and directional differences of each feature on the model prediction are especially noteworthy. For instance, main engine power and gross tonnage exhibit a significant linear effect over a certain range of values. Specifically, as the main engine power and gross tonnage increase, the predicted impact on accident severity gradually shifts from positive to negative, indicating that ships with higher engine power and larger tonnage are more effective in avoiding serious accidents in rough sea conditions compared to those with lower power and smaller tonnage. This phenomenon suggests that ships with higher power and greater tonnage possess better stability and risk resistance in emergency situations, thereby reducing the likelihood of accidents [12]. Moreover, the deep draft ratio and time display a more complex non-linear relationship, making it difficult to generalize their effects on accident severity as simply positive or negative. The deep draft ratio, which reflects ship stability and adaptability to sea conditions, may be linked to accident severity under specific conditions, while the different voyage times may correlate with increased operational fatigue and sudden risk, exhibiting non-linear sensitivity [80].

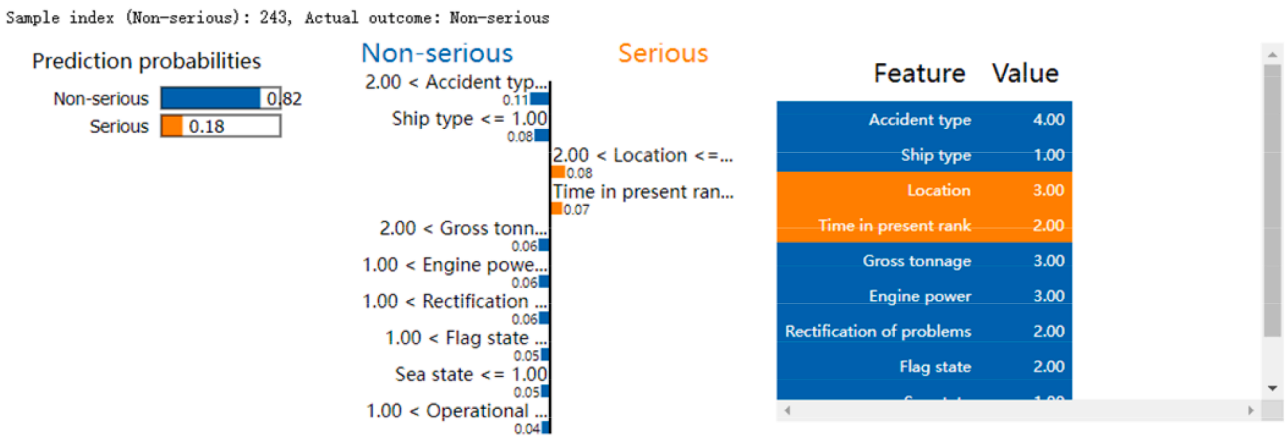
In summary, this study not only quantifies the importance of each feature in predicting maritime accident severity through the application of SHAP values but also enhances the understanding of the model's decision logic. The visualization results of the SHAP values clearly illustrate both the overall and individual effects of feature contributions, revealing the interactions between key features and their non-linear relationships.

4.2.2. Local analysis with LIME

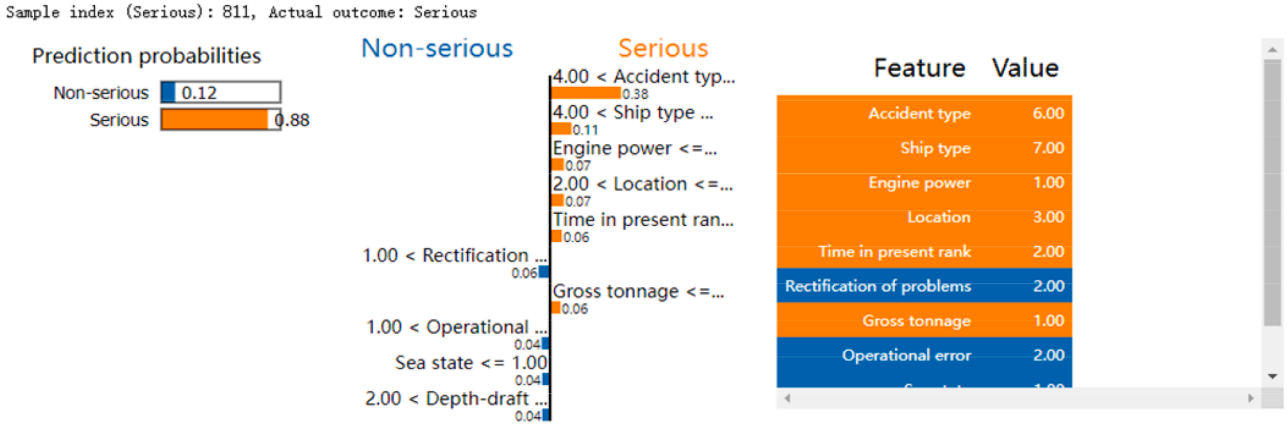
This section demonstrates the use of LIME for localized interpretation of individual predictions in maritime accident severity models, allowing for a detailed analysis of feature contributions at the instance level. The aim of this study is to provide insights into the application of the LIME technique for local interpretation for individual samples in the constructed models. LIME, as an advanced interpretability tool, uncovers the underlying drivers of complex model decisions by generating a local linear model for each individual prediction [36,81]. To effectively demonstrate the practical application of this technique, this study randomly selected a serious accident and a non-serious one as samples for analysis. Through this comparative analysis, this study not only identifies the key risk factors affecting accident severity but also reveals the differences in the importance of these factors and their interactions in various accident scenarios. The results of this partial analysis offer valuable insights into understanding the model's behaviour and decision-making basis in specific contexts, thus enhancing model transparency and interpretability.

Fig. 11 illustrates the ML interpretation of accident severity prediction for two random samples using the LIME technique. Fig. 11 is divided into three main sections: the predicted probability, the feature interpretation, and the table of normalized values. Specifically, Sample 1 is predicted to be a non- serious accident with a predicted probability of 82%, and Sample 2 is predicted to be a serious accident with a predicted probability of 88%. To more clearly present the positive and negative contributions of each feature to the prediction results and their specific values, Fig. 12 is presented. In Fig. 12, orange bars represent the positive influence on the prediction of the sample as a serious accident, while blue bars represent the negative influence on the prediction of the sample as a serious accident.

As shown in Fig. 12, for the non-serious accident sample (Sample 1), only two features—the location of the navigational waters and the crew's time in present rank —contributed positively to the prediction of it being a serious accident. All other features played a positive role in predicting it as a non-serious accident. This finding suggests that, although Sample 1 is overall predicted as a non-serious accident, the navigational waters and time in rank under specific conditions may still influence its risk assessment results, highlighting the potential significance of environmental factors and crew rank in accident occurrences. Specifically, Sample 1 occurred in coastal waters, and the crew's time in present rank ranged from one to five years. This combination suggests that accident risk assessments may be influenced by a mix of geographical and human factors under specific operational circumstances and experience levels. Coastal waters often present more complex navigational challenges and environmental variables, while the relative inexperience of crew members in their rank may lead to underestimating risks and responding untimely [46,82]. Such conditions can exacerbate accident severity, particularly in dynamic climatic conditions and variable water environments, underscoring the strong link between crew operational competence and accident outcomes. Integrating environmental and human factors into maritime safety management is therefore crucial to mitigate accident risks and enhance emergency response efficiency. In contrast, Sample 2, categorized as a serious accident, showed that several key features significantly increased its probability of being predicted as severe. However, factors such as timely rectification of problems, relatively smooth sea state (rating 0–3), the absence of flags of convenience, and the lack of evident operational error positively mitigated the probability of a severe prediction. This aligns with the actual scenario, further validating the model's accuracy in identifying complex accident scenarios. Notably, the timely resolution of issues and lower sea state risk demonstrate that effective safety management and stable navigational conditions can help reduce accident severity even in high-risk scenarios [83]. These findings highlight how LIME analyses reveal the granular impacts of individual characteristics on accident predictions and provide insights into the interplay of risk factors across different scenarios. This approach offers



(a) Sample 1-Non-serious



(b) Sample 2-Serious

Fig. 11. LIME analysis: non-serious versus serious characterisation impacts.

quantitative data for accurate predictions and the formulation of targeted prevention strategies.

4.2.3. *Validation of LIME' s local explanation reliability using SHAP*

To further verify the reliability of LIME in interpreting local predictions, this study selects a serious accident sample (Sample 2) for in-depth analysis, incorporating SHAP to conduct a multi-level, multi-perspective consistency test. As illustrated in Fig. 13, SHAP provides a detailed visualization of the feature influences in Sample 2, revealing the role of individual features in predicting accident severity and their respective positive or negative effects. To enhance the clarity of results, Fig. 14 uses "+" and "−" symbols to denote positive and negative impacts on the prediction, respectively. The analysis demonstrates a high degree of consistency between LIME and SHAP regarding the positive and negative impacts of features. However, some differences are observed. For positive influence features, LIME highlights "time in present crew rank" more prominently than SHAP, underscoring the potential contribution of crew experience to risk prediction and demonstrating LIME's fine-grained capacity to identify individual factors in local interpretations. Conversely, for negative influence features, LIME includes "sea state," "flag of convenience," and "operational error" more comprehensively than SHAP, enhancing the model's interpretability and applicability to real-world scenarios. Meanwhile, SHAP uniquely identifies the feature "depth-draft ratio", which may be attributed to its inherent sensitivity to global feature importance, even when applied to local explanations. This suggests that SHAP's individual case interpretations can still be influenced by a feature's overall contribution across the dataset, potentially introducing a degree of bias. Notably, in Section 4.2.1, the SHAP-based global analysis highlights 'draught depth ratio' as a highly influential feature, further corroborating this perspective.

This comparison highlights the significant advantages of LIME in local prediction, offering a unique perspective and a reliable foundation for analysing feature contributions in complex prediction tasks. By accurately identifying and deconstructing influencing factors, LIME enhances the depth and adaptability of local interpretations for individual samples. In contrast, SHAP provides a globally consistent view of feature importance, while LIME's flexibility in sample-specific interpretation further strengthens the credibility and interpretive value of its analysis. Overall, the comparison presents a multidimensional perspective on model interpretation, demonstrating the complementary nature of LIME and SHAP in complex prediction tasks. This synergy provides a robust theoretical foundation for ensuring the stability and consistency of local interpretations, further supporting the applicability and reliability of these models in predicting accident severity.

5. Implications

This study contributes to maritime accident severity prediction by addressing key challenges in machine learning applications, particularly in unbalanced datasets and complex predictive modelling. The research systematically advances knowledge in data processing, model selection,

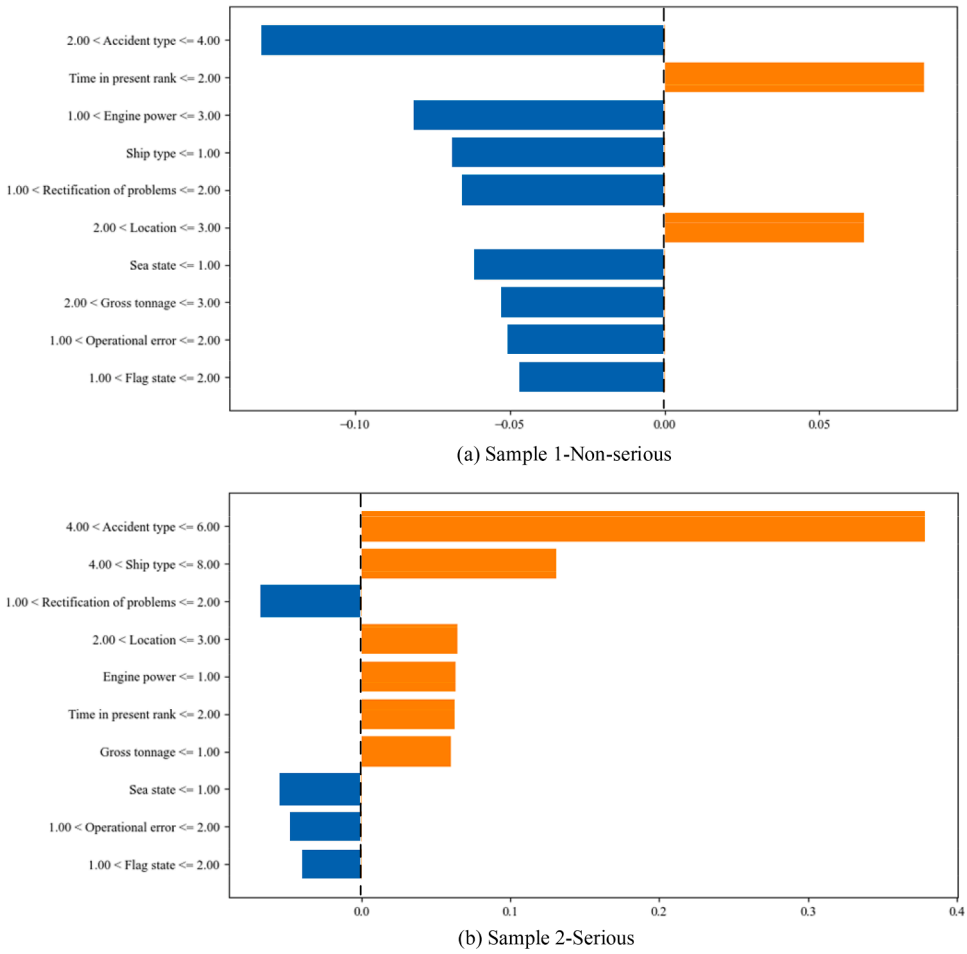


Fig. 12. LIME explanation for non-serious and serious Outcome (for the two samples in Fig. 11).

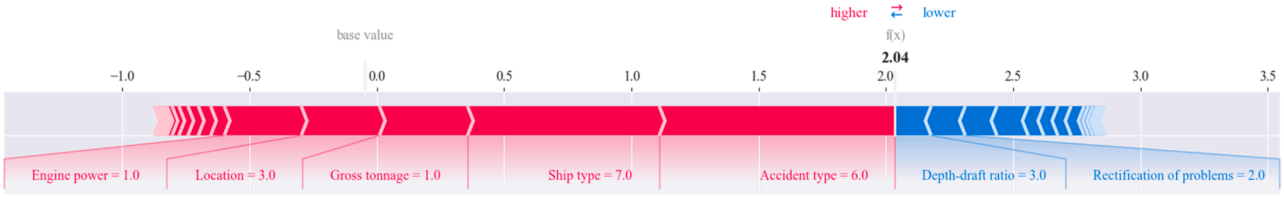


Fig. 13. The local interpretation consistency validation plot for Sample 2.

and interpretability, with significant implications for both theoretical understanding and practical safety management.

Firstly, the prediction framework based on data balancing and ML developed in this study provides a more accurate risk warning tool for maritime safety management. By combining the optimal balancing method with the ML model, the RandomOverSampler-CatBoost combination efficiently predicts several types of accidents. The framework not only significantly enhances the accuracy of accident severity prediction but also provides valuable technical support to decision-makers, helping to mitigate the potential threat of maritime accidents to life and property. Specifically, in the context of unbalanced data, the balanced strategy employed in this study substantially improves the detection rate of rare serious accidents, enabling managers to take more targeted preventive measures against high-risk events, thereby optimizing resource allocation and maximizing safety outcomes.

Secondly, this study identifies key variables that influence the severity of maritime accidents, providing a solid foundation for developing precise, data-driven safety measures. Through multi-perspective feature interpretation using SHAP and LIME, this study clarifies the extent to which important features, such as accident type, ship type, main engine power, and gross tonnage, influence accident prediction. This information can help maritime safety authorities promote precise prevention policies. For example, ship type and main engine power may present higher accident risks in specific waters, and policymakers can implement targeted regulatory measures in high-risk areas to prevent elevated accident rates of specific ship with certain type or main engine power. Additionally, safety inspections and training can be strengthened for smaller gross tonnage vessels, particularly in adverse weather and challenging environments. This data-driven strategy not only advances the science of safety management but also reduces accident frequency and effectively lowers the risk of serious accidents.

Thirdly, the explanatory analyses of SHAP and LIME provide reliable technical support for identifying the causes of accidents in specific contexts. The global analysis using SHAP helps identify the most influential factors in accident prediction, offering risk managers a macroscopic view and supporting the development of comprehensive

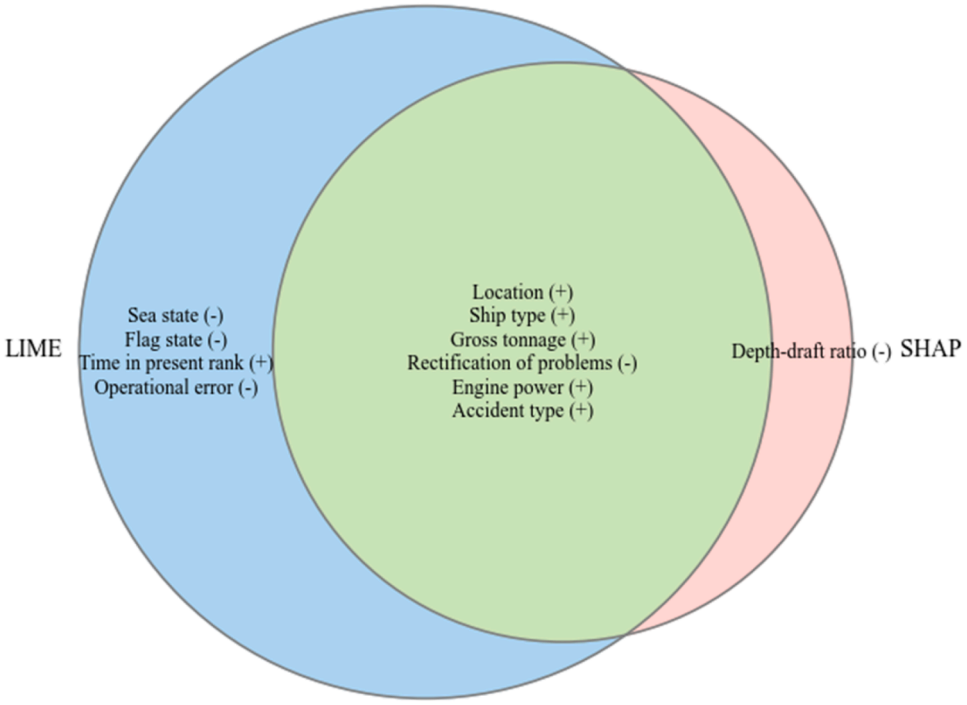


Fig. 14. Comparative analysis of feature influences using LIME and SHAP for Sample 2.

prevention programs. In contrast, the local analysis using LIME reveals the contribution of specific factors in individual accident scenarios, enabling managers to better understand the sources of risk in each case. For instance, in a particular accident prediction scenario, this study found that variables such as operating time, crew experience, and navigational waters significantly impacted the prediction results. By localizing the interpretation of these factors, maritime stakeholders can pinpoint high-risk variables under specific operating conditions, thereby targeting risk interventions to reduce potential safety hazards.

Finally, the multi-level explanatory framework proposed in this study offers practical guidance for developing maritime accident prediction systems. The combination of SHAP and LIME enhances the transparency of decision-making in complex models and provides clear guidelines for technicians developing accident prediction systems, making these systems more adaptable and flexible in handling key features. Furthermore, the reliability validation of this framework in practical applications serves as a baseline for risk prediction systems in other high-risk sectors (e.g., aviation, transportation), promoting the design and application of intelligent safety management systems across industries, thus enhancing overall industry safety.

6. Conclusions

This study presents an integrated framework for predicting the severity of maritime accidents using multiple data balancing techniques and ML models, coupled with interpreting the model results from global and local perspectives.

Firstly, the integrated prediction framework based on multiple data balancing techniques significantly improves the prediction of maritime accident severity. Through comparison of six data balancing techniques, it is revealed that methods such as RandomOverSampler and SMOTE-Tomek effectively enhanced the identification of rare event types. The application of these techniques markedly improved classification accuracy on unbalanced datasets, providing more reliable technical support for maritime accident prediction. Additionally, the advanced ML models, such as CatBoost, LightGBM, and XGBoost, used in this study ensure preferable performances across various metrics through hyperparameter tuning. The RandomOverSampler-CatBoost combination is identified as the best-performing model combination, demonstrating superior adaptability and predictive accuracy in handling complex datasets.

Secondly, the model's interpretability is another significant contribution of this study. The global feature contribution analysis via SHAP clarifies the impact of key factors such as accident type, ship type, main engine power, and gross tonnage on prediction results, while local interpretation for individual samples through LIME enhances understanding of sample-specific predictions. The combination of both methods offers a dual global and local perspective, increasing model transparency and providing valuable insights for policymaker decision-making. The explanatory analytical framework not only deepens the understanding of accident severity prediction but also provides data-driven support for targeted safety management measures.

Finally, the prediction framework demonstrates broad practical applicability. The feature importance analysis, based on the multi-level explanatory framework, offers systematic support for developing accident prevention and emergency management strategies, particularly valuable in high-risk maritime scenarios. In summary, this study provides a multidimensional, interpretable, and accurate solution for maritime accident severity prediction, advances the application of data imbalance-aware prediction models in the maritime domain, and offers robust theoretical and technical support for future safety management initiatives.

Despite the promising results of this study, there are still limitations in terms of research methodology and data resources, which provide directions for future research. Firstly, the regional scope of the data sources and the limited time span of this study may be addressed in future to enhance the generalizability of the model. Future studies could incorporate more international and diverse maritime accident data (i.e. the IMO's GISIS) with a longer time frame. Secondly, while this study employed rigorous hold-out and cross-validation techniques, a future validation would involve testing the trained models against a completely independent, unseen dataset. Such an external validation would provide the most stringent assessment of the framework's real-world predictive power and resilience across disparate operational contexts. Finally,

while this study developed an integrated framework with multiple data balancing techniques and ML models, some methods may be computationally inefficient when applied to large-scale datasets. Future research could explore the use of ensemble learning or distributed computing techniques to improve processing efficiency in big data environments.

CRediT authorship contribution statement

Wenjie Cao: Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. **Xinjian Wang:** Writing – original draft, Visualization, Supervision, Methodology, Funding acquisition, Conceptualization. **Yuanjun Feng:** Writing – original draft, Validation, Resources, Methodology, Investigation, Formal analysis. **Jingen Zhou:** Writing – original draft, Visualization, Resources, Methodology, Formal analysis. **Zaili Yang:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Methodology, Funding acquisition, Formal analysis.

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.

Acknowledgements

This work is financially supported by the National Natural Science Foundation of China [Grant No. 52101399], and the MarRI-UK Core Project [Grant Agreement No. MarRI-UK_CORE_2024_02]. This work was also supported by the European Union Horizon 2020 European Research Council Consolidator Grant program [TRUST Grant No. 864724].

Appendix A: Supplementary data

The feature lists to this article can be found at the supplementary file.
Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ress.2025.111648.

Appendix B: Hyperparameter optimization and model performance metrics

To ensure methodological consistency and equitable benchmarking across all model–balancing configurations, a unified grid search strategy was applied to each of the eight machine learning algorithms evaluated in this study. Hyperparameter optimisation was conducted separately for each of the 48 combinations arising from pairing six data balancing techniques with eight machine learning models. Importantly, the hyperparameter search space for each model remained fixed across all data balancing methods, thereby preserving consistency in comparative evaluation. For LR, the regularisation strength parameter C was explored over the range [0.001, 0.01, 0.1, 1, 10], with penalty types selected from ["l1", "l2"], and the solver fixed as "liblinear" to support *l1* regularisation. For the DT model, the search grid comprised *max_depth* ∈ [5, 10, 20, None], *min_samples_split* ∈ [2, 5, 10], and *min_samples_leaf* ∈ [1, 2, 4]. For ensemble-based models, RF and ET adopted identical search spaces defined by *n_estimators* ∈ [100, 200, 500], *max_depth* ∈ [10, 20, None], *min_samples_split* ∈ [2, 5, 10], and *min_samples_leaf* ∈ [1, 2, 4]. For gradient boosting algorithms, LightGBM was tuned using *n_estimators* ∈ [100, 200, 500], *learning_rate* ∈ [0.01, 0.05, 0.1], *max_depth* ∈ [10, 20, 30], and *num_leaves* ∈ [31, 50, 100], while XGBoost employed *n_estimators* ∈ [100, 200, 500], *learning_rate* ∈ [0.01, 0.05, 0.1], *max_depth* ∈ [6, 10, 15], and *colsample_bytree* ∈ [0.7, 0.8, 1.0]. The CatBoost algorithm was optimised over *iterations* ∈ [500, 1000], *learning_rate* ∈ [0.01, 0.05, 0.1], and *depth* ∈ [6, 10, 15]. Finally, the NN model was configured using *hidden_layer_sizes* ∈ [(100,), (50, 50), (100, 50)], *learning_rate_init* ∈ [0.001, 0.01, 0.1], and *max_iter* ∈ [500, 1000]. For each model–balancing combination, a grid search with five-fold cross-validation was performed, and the configuration that yielded the highest validation performance was selected as optimal. The results of this tuning process, including the final selected hyperparameters for all 48 model–balancing combinations, are detailed in Table B1. This rigorous and standardised optimisation protocol enhances the credibility, fairness, and reproducibility of the model evaluation presented in this study.

Table B1
Summary of hyperparameter optimisation results for ML models.

Data balancing techniques	Model	Selected setting
SMOTE	LR	[C: 0.1, penalty: "l2", solver: "liblinear"]
	DT	[max_depth: 5, min_samples_leaf: 4, min_samples_split: 2]
	RF	[max_depth: None, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 500]
	ET	[max_depth: None, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 500]
	NN	[hidden_layer_sizes: (100, 50), learning_rate_init: 0.001, max_iter: 1000]
	LightGBM	[learning_rate: 0.1, max_depth: 10, n_estimators: 200, num_leaves: 50]
	XGBoost	[colsample_bytree: 0.7, learning_rate: 0.1, max_depth: 6, n_estimators: 100]
	CatBoost	[depth: 10, iterations: 1000, learning_rate: 0.01]
BorderlineSMOTE	LR	[C: 0.1, penalty: "l2", solver: "liblinear"]
	DT	[max_depth: 5, min_samples_leaf: 4, min_samples_split: 2]
	RF	[max_depth: 20, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 500]
	ET	[max_depth: None, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 200]
	NN	[hidden_layer_sizes: (100, 50), learning_rate_init: 0.001, max_iter: 1000]
	LightGBM	[learning_rate: 0.1, max_depth: 20, n_estimators: 200, num_leaves: 50]
	XGBoost	[colsample_bytree: 0.8, learning_rate: 0.1, max_depth: 10, n_estimators: 100]
	CatBoost	[depth: 10, iterations: 1000, learning_rate: 0.05]
KMeansSMOTE	LR	[C: 0.1, penalty: "l2", solver: "liblinear"]
	DT	[max_depth: 5, min_samples_leaf: 4, min_samples_split: 5]
	RF	[max_depth: 20, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 500]
	ET	[max_depth: 20, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 200]
	NN	[hidden_layer_sizes: (100,), learning_rate_init: 0.001, max_iter: 1000]
	LightGBM	[learning_rate: 0.05, max_depth: 20, n_estimators: 100, num_leaves: 50]
	XGBoost	[colsample_bytree: 0.7, learning_rate: 0.01, max_depth: 15, n_estimators: 200]

(*continued on next page*)

Table B1 (*continued*)

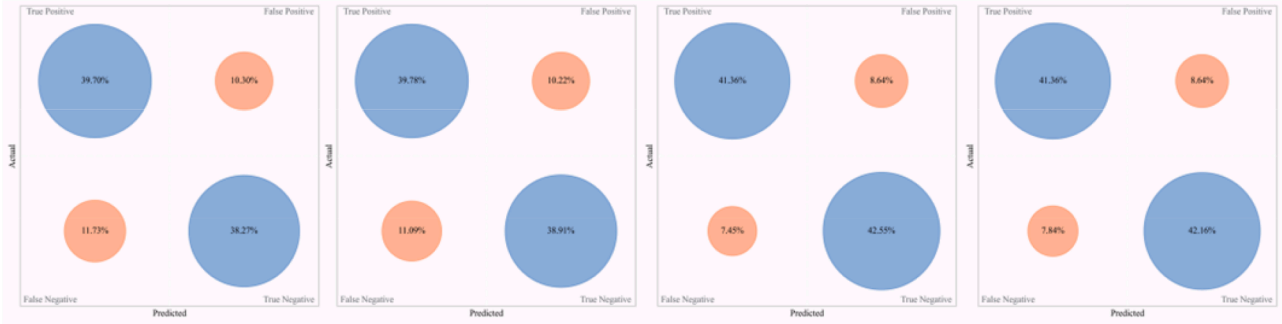
Data balancing techniques	Model	Selected setting
	CatBoost	[depth: 10, iterations: 1000, learning_rate: 0.01]
RandomOverSampler	LR	[C: 0.1, penalty: "l2", solver: "liblinear"]
	DT	[max_depth: None, min_samples_leaf: 4, min_samples_split: 10]
	RF	[max_depth: 20, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 500]
	ET	[max_depth: None, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 100]
	NN	[hidden_layer_sizes: (100, 50), learning_rate_init: 0.001, max_iter: 1000]
	LightGBM	[learning_rate: 0.1, max_depth: 20, n_estimators: 200, num_leaves: 50]
	XGBoost	[colsample_bytree: 0.7, learning_rate: 0.1, max_depth: 10, n_estimators: 200]
	CatBoost	[depth: 15, iterations: 1000, learning_rate: 0.01]
SMOTETomek	LR	[C: 0.1, penalty: "l2", solver: "liblinear"]
	DT	[max_depth: 5, min_samples_leaf: 4, min_samples_split: 10]
	RF	[max_depth: 20, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 200]
	ET	[max_depth: 20, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 200]
	NN	[hidden_layer_sizes: (100,), learning_rate_init: 0.001, max_iter: 1000]
	LightGBM	[learning_rate: 0.05, max_depth: 20, n_estimators: 500, num_leaves: 50]
	XGBoost	[colsample_bytree: 0.8, learning_rate: 0.1, max_depth: 15, n_estimators: 100]
	CatBoost	[depth: 10, iterations: 1000, learning_rate: 0.05]
SVMSMOTE	LR	[C: 0.1, penalty: "l2", solver: "liblinear"]
	DT	[max_depth: 5, min_samples_leaf: 4, min_samples_split: 10]
	RF	[max_depth: 20, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 200]
	ET	[max_depth: 20, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 500]
	NN	[hidden_layer_sizes: (100, 50), learning_rate_init: 0.001, max_iter: 1000]
	LightGBM	[learning_rate: 0.1, max_depth: 10, n_estimators: 200, num_leaves: 50]
	XGBoost	[colsample_bytree: 0.8, learning_rate: 0.05, max_depth: 10, n_estimators: 200]
	CatBoost	[depth: 10, iterations: 1000, learning_rate: 0.01]

Table B2
The result of the performance metrics of various models without data balancing techniques.

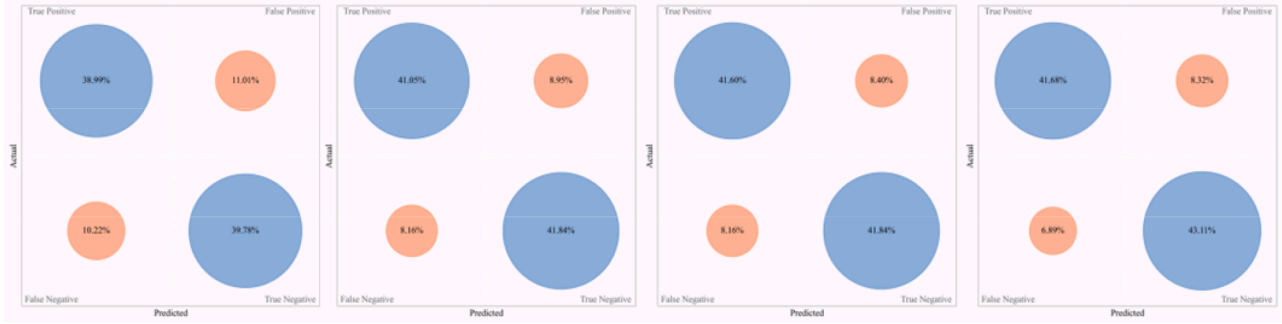
Model	Accuracy	Precision	Recall	F1-score	ROC AUC
LR	0.7529	0.7312	0.6355	0.6800	0.7942
DT	0.7529	0.7009	0.7009	0.7009	0.7452
RF	0.7699	0.7604	0.6822	0.7192	0.8421
ET	0.7954	0.8068	0.6636	0.7282	0.8469
NN	0.7529	0.6807	0.7570	0.7168	0.8134
LightGBM	0.7954	0.7547	0.7477	0.7512	0.8332
XGBoost	0.7735	0.7232	0.7570	0.7397	0.8327
CatBoost	0.7759	0.7500	0.7009	0.7246	0.8404

Note: All evaluation metrics (Accuracy, Precision, Recall, F1-score, and ROC AUC) are based on five-fold cross-validation.

Appendix C: Confusion matrix

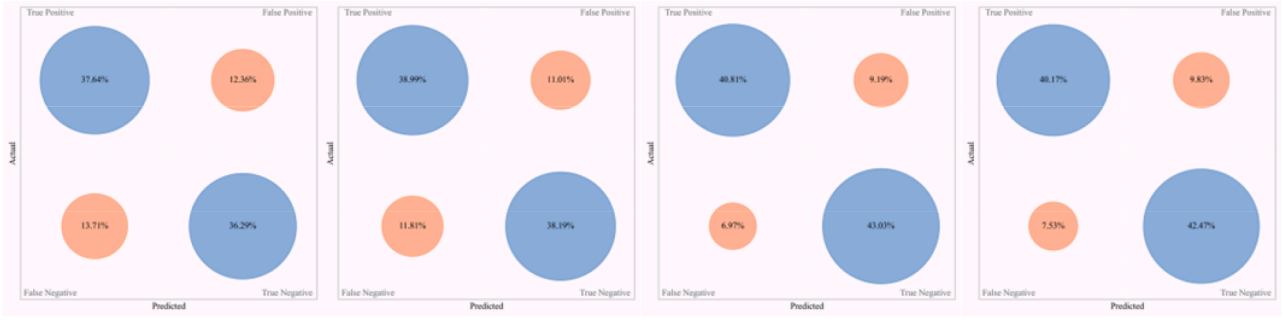


(a) SMOTE-LR (b) SMOTE-DT (c) SMOTE-RF (d) SMOTE-ET



(e) SMOTE-NN (f) SMOTE-LightGBM (g) SMOTE-XGBoost (h) SMOTE-CatBoost

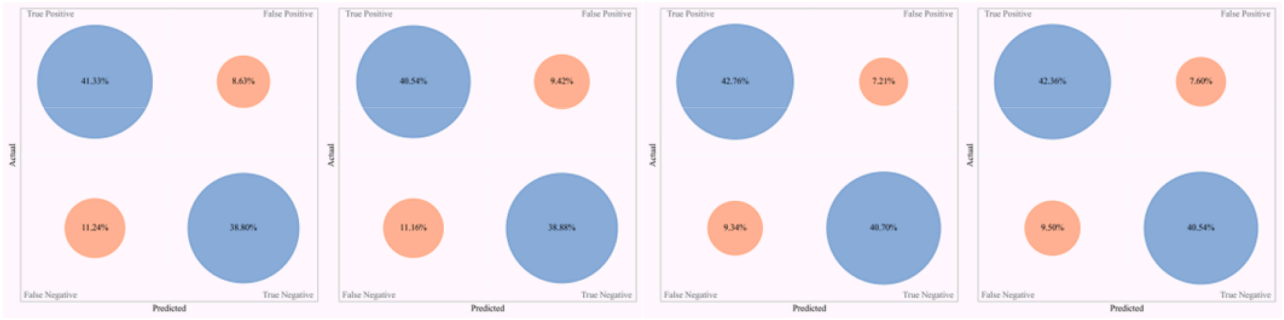
Fig. C1. The result of confusion matrix for SMOTE.



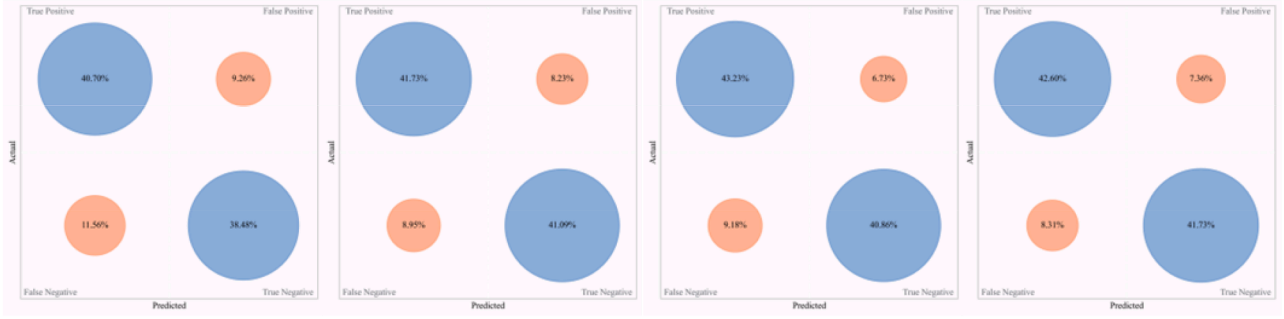
(a) BorderlineSMOTE-LR (b) BorderlineSMOTE-DT (c) BorderlineSMOTE-RF (d) BorderlineSMOTE-ET

(e) BorderlineSMOTE-NN (f) BorderlineSMOTE-LightGBM (g) BorderlineSMOTE-XGBoost (h) BorderlineSMOTE-CatBoost

Fig. C2. The result of confusion matrix for BorderlineSMOTE.

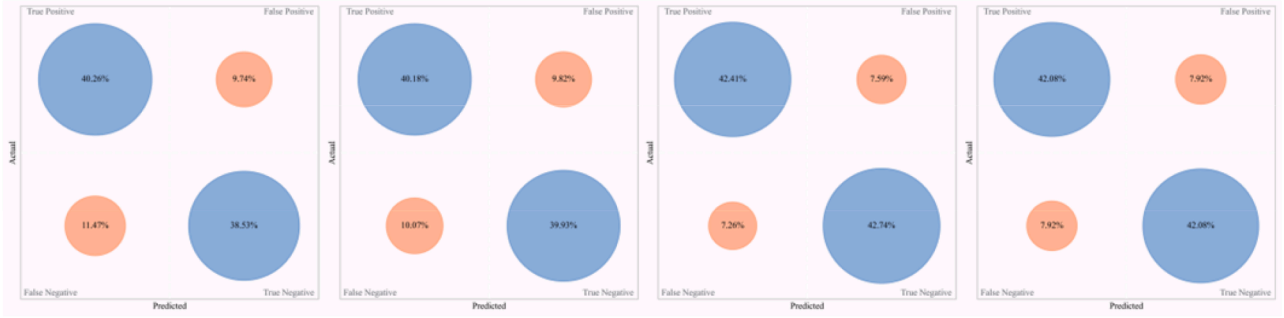


(a) KMeansSMOTE-LR (b) KMeansSMOTE-DT (c) KMeansSMOTE-RF (d) KMeansSMOTE-ET

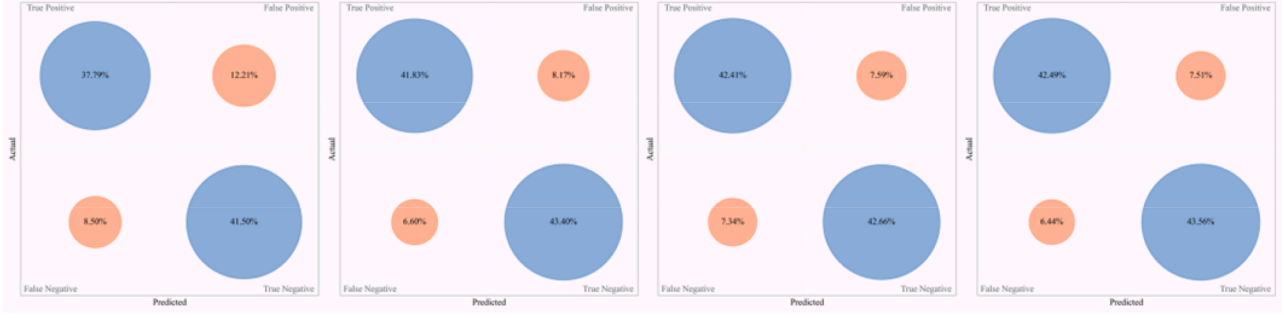


(e) KMeansSMOTE-NN (f) KMeansSMOTE-LightGBM (g) KMeansSMOTE-XGBoost (h) KMeansSMOTE-CatBoost

Fig. C3. The result of confusion matrix for KMeansSMOTE.

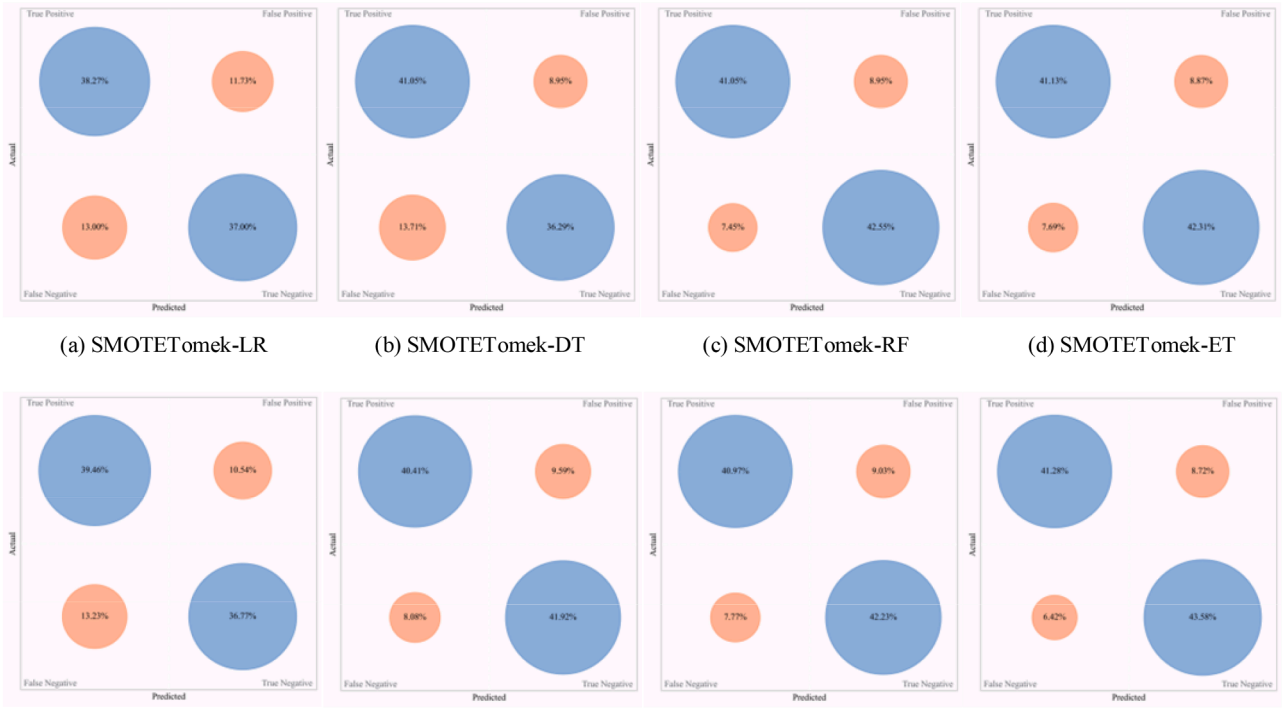


(a) SMOTETomek-LR (b) SMOTETomek-DT (c) SMOTETomek-RF (d) SMOTETomek-ET



(e) SMOTETomek-NN (f) SMOTETomek-LightGBM (g) SMOTETomek-XGBoost (h) SMOTETomek-CatBoost

Fig. C4. The result of confusion matrix for SMOTETomek.



(a) SMOTETomek-LR (b) SMOTETomek-DT (c) SMOTETomek-RF (d) SMOTETomek-ET

(e) SMOTETomek-NN (f) SMOTETomek-LightGBM (g) SMOTETomek-XGBoost (h) SMOTETomek-CatBoost

Fig. C5. The result of confusion matrix for SVMSMOTE.

Appendix D: Supplementary material

The source code is publicly available at: https://github.com/AdvMarTech/BalancedMaritimeAccidentXAI.

Data availability

Data will be made available on request.

References

[1] Cao WJ, Wang XJ, Li J, Zhang ZW, Cao YH, Feng YW. A novel integrated method for heterogeneity analysis of marine accidents involving different ship types. Ocean Eng 2024;312:119295. https://doi.org/10.1016/j.oceaneng.2024.119295.

[2] Feng YW, Wang XJ, Chen QL, Yang ZL, Wang J, Li HH, Xia GQ, Liu ZJ. Prediction of the severity of marine accidents using improved machine learning. Transp Res E-Logist Transp Rev 2024;188:103647. https://doi.org/10.1016/j.tre.2024.103647.

[3] Wang X, Cao W, Li T, Feng Y, Uğurlu Ö, Wang J. An integrated multidimensional model for heterogeneity analysis of maritime accidents during different watchkeeping periods. Ocean Coast Manag 2025;264:107625. https://doi.org/10.1016/j.ocecoaman.2025.107625.

[4] European Maritime Safety Agency, 2024. Annual overview of marine casualties and incidents 2024. http://www.emsa.europa.eu/news-a-press-centre/external-news/item/3734-annual-overview-of-marine-casualties-and-incidents-2024.htm.

[5] Zhang Z, Liu Z, Zhou Z, Wang X, Majumdar A, Cao Y, Yang Z. Time prediction of human evacuation from passenger ships based on machine learning methods. Transport Res Part A: Policy Pract 2025;200:104644. https://doi.org/10.1016/j.tra.2025.104644.

[6] Tao J, Liu Z, Wang X, Cao Y, Matthews C, Yang Z. Advanced modelling for team collaborative decision making analysis on maritime autonomous surface ships using team cognitive work analysis. Regional Stud in Marine Sci 2025;90:104477. https://doi.org/10.1016/j.rsma.2025.104477.

[7] Brandt P, Munim ZH, Chaal M, Kang HS. Maritime accident risk prediction integrating weather data using machine learning. Transp Res D-Transp Environ 2024;136:104388. https://doi.org/10.1016/j.trd.2024.104388.

[8] Cao Y, Xin X, Wang X, Wang J, Yang Z. Multi-objective resilience-oriented optimisation for the global container shipping network against cascading failures. Transp Res A-Policy Pract 2025;200:104659. https://doi.org/10.1016/j.tra.2025.104659.

[9] Cao Y, Iulia M, Majumdar A, Feng Y, Xin X, Wang X, Wang H, Yang Z. Investigation of the risk influential factors of maritime accidents: a novel topology and robustness analytical framework. Reliab Eng Syst Saf 2025;254:110636. https://doi.org/10.1016/j.ress.2024.110636.

[10] Munim ZH, Sorli MA, Kim H, Alon I. Predicting maritime accident risk using automated machine learning. Reliab Eng Syst Saf 2024;248:110148. https://doi.org/10.1016/j.ress.2024.110148.

[11] Wang JH, Zhou Y, Zhuang L, Shi L, Zhang SG. A model of maritime accidents prediction based on multi-factor time series analysis. J Mar Eng Technol 2023;22(3):153–65. https://doi.org/10.1080/20464177.2023.2167269.

[12] Cao YH, Wang XJ, Wang YH, Fan SQ, Wang HX, Yang ZL, Liu ZJ, Wang J, Shi RJ. Analysis of factors affecting the severity of marine accidents using a data-driven Bayesian network. Ocean Eng 2023;269:113563. https://doi.org/10.1016/j.oceaneng.2022.113563.

[13] Douzas G, Bacao F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE. Inf Sci 2019;501:118–35. https://doi.org/10.1016/j.ins.2019.06.007.

[14] Ma L, Ma X, Du Q, Zhang R. Investigation of the severity of maritime accidents considering the interaction between human factors and operating conditions: a case study on collision accidents in China. Reliab Eng Syst Saf 2025;265:111533. https://doi.org/10.1016/j.ress.2025.111533.

[15] Kandel R, Baroud H. A data-driven risk assessment of Arctic maritime incidents: using machine learning to predict incident types and identify risk factors. Reliab Eng Syst Saf 2024;243:109779. https://doi.org/10.1016/j.ress.2023.109779.

[16] Ung, S.-T.Z., 2007. The development of safety and security assessment techniques and their application to port operations.

[17] Akyuz E, Celik M, Cebi S. A phase of comprehensive research to determine marine-specific EPC values in human error assessment and reduction technique. Saf Sci 2016;87:63–75. https://doi.org/10.1016/j.ssci.2016.03.013.

[18] Yu J, Zhao J, Wang X, Cao Y. Maritime occupational accidents analysis: A data-driven Bayesian network approach. Ocean Coast Manage 2025;269:107785. https://doi.org/10.1016/j.ocecoaman.2025.107785.

[19] Aydin M, Sezer SI, Akyuz E, Gardoni P. Improved Z-number based Bayesian network modelling to predict cyber-attack risk for maritime autonomous surface ship (MASS). App Soft Comput 2025;180:113416. https://doi.org/10.1016/j.asoc.2025.113416.

[20] Kong D, Lin Z, Li W, He W. Development of an improved Bayesian network method for maritime accident safety assessment based on multiscale scenario analysis theory. Reliab Eng Syst Saf 2024;251:110344. https://doi.org/10.1016/j.ress.2024.110344.

[21] Li H, Çelik C, Bashir M, Zou L, Yang Z. Incorporation of a global perspective into data-driven analysis of maritime collision accident risk. Reliab Eng Syst Saf 2024; 249:110187. https://doi.org/10.1016/j.ress.2024.110187.
[22] Yang Y, Shao ZP, Hu Y, Mei Q, Pan JC, Song RX, Wang P. Geographical spatial analysis and risk prediction based on machine learning for maritime traffic accidents: a case study of Fujian sea area. Ocean Eng 2022;266:113106. https://doi.org/10.1016/j.oceaneng.2022.113106.
[23] Kim G, Lim S. Development of an interpretable maritime accident prediction system using machine learning techniques. IEEE Access 2022;10:41313–29. https://doi.org/10.1109/ACCESS.2022.3168302.
[24] Atak Ü, Arslanoglu Y. Machine learning methods for predicting marine port accidents: a case study in container terminal. Sh Offshore Struct 2022;17(11): 2480–7. https://doi.org/10.1080/17445302.2021.2003067.
[25] Rao CJ, Wei X, Xiao XP, Shi Y, Goh M. Oversampling method via adaptive double weights and Gaussian kernel function for the transformation of unbalanced data in risk assessment of cardiovascular disease. Inf Sci 2024;665:120410. https://doi.org/10.1016/j.ins.2024.120410.
[26] Rao CJ, Huang QF, Chen L, Goh M, Hu Z. Forecasting the carbon emissions in Hubei Province under the background of carbon neutrality: a novel STIRPAT extended model with ridge regression and scenario analysis. Environ Sci Pollut Res 2023;30(20):57460–80. https://doi.org/10.1007/s11356-023-26599-w.
[27] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57. https://doi.org/10.1613/jair.953.
[28] Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang DS, Zhang XP, Huang GB, editors. Proceedings of the advances in intelligent computing, part 1; 2005. p. 878–87. https://doi.org/10.1007/11538059_91.
[29] Wu J, Zhao Z, Sun C, Yan R, Chen X. Learning from class-imbalanced data with a model-agnostic framework for machine intelligent diagnosis. Reliab Eng Syst Saf 2021;216:107934. https://doi.org/10.1016/j.ress.2021.107934.
[30] Ding Y, Jia M, Zhuang J, Cao Y, Zhao X, Lee C-G. Deep imbalanced domain adaptation for transfer learning fault diagnosis of bearings under multiple working conditions. Reliab Eng Syst Saf 2023;230:108890. https://doi.org/10.1016/j.ress.2022.108890.
[31] Wu Z, Xu R, Luo Y, Shao H. A holistic semi-supervised method for imbalanced fault diagnosis of rotational machinery with out-of-distribution samples. Reliab Eng Syst Saf 2024;250:110297. https://doi.org/10.1016/j.ress.2024.110297.
[32] Abdulsadig RS, Rodriguez-Villegas E. A comparative study in class imbalance mitigation when working with physiological signals. Front Digit Health 2024;6: 1377165. https://doi.org/10.3389/fdgth.2024.1377165. -1377165.
[33] Korkmaz A, Bulut S, Talan T, Kosunalp S, Iliev T. Enhancing firewall packet classification through artificial neural networks and synthetic minority over-sampling technique: an innovative approach with evaluative comparison. Appl Sci 2024;14(16):7426. https://doi.org/10.3390/app14167426.
[34] Abou Elassad DE, Abou Elassad ZE, Ed-Dahbi AM, El Meslouhi O, Kardouchi M, Akhloufi M. An advanced accident avoidance system based on imbalance-control ensemble and deep learning fusion design. J Intell Fuzzy Syst 2024;46(2):3659–76. https://doi.org/10.3233/JIFS-232446.
[35] Abdulrashid I, Farahani RZ, Mammadov S, Khalafalla M, Chiang WC. Explainable artificial intelligence in transport logistics: risk analysis for road accidents. Transp Res E-Logist Transp Rev 2024;186:103563. https://doi.org/10.1016/j.tre.2024.103563.
[36] Xie CD, Hu JW, Vasdravellis G, Wang XT, Cheng SB. Explainable AI model for predicting equivalent viscous damping in dual frame-wall resilient system. J Build Eng 2024;96:110564. https://doi.org/10.1016/j.jobe.2024.110564.
[37] Chimatapu R, Hagras H, Starkey A, Owusu G. Explainable AI and Fuzzy Logic Systems, Theory and Practice of Natural Computing: 7th International Conference, TPNC 2018, Dublin, Ireland, December 12-14, 2018. Dublin, Ireland: Springer-Verlag; 2018. p. 3–20.
[38] Yuan C, Li Y, Huang HL, Wang SQ, Sun ZH, Wang HG. Application of explainable machine learning for real-time safety analysis toward a connected vehicle environment. Accid Anal Prev 2022;171:106681. https://doi.org/10.1016/j.aap.2022.106681.
[39] Vimbi V, Shaffi N, Mahmud M. Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection. Brain Inform 2024;11(1):10. https://doi.org/10.1186/s40708-024-00222-1. -10.
[40] Zhang C, Zou X, Lin C. Fusing XGBoost and SHAP models for maritime accident prediction and causality interpretability analysis. J Mar Sci Eng 2022;10(8):1154. https://doi.org/10.3390/jmse10081154.
[41] He J, Hao Y, Wang XQ. An interpretable aid decision-making model for flag state control ship detention based on SMOTE and XGBoost. J Mar Sci Eng 2021;9(2): 156. https://doi.org/10.3390/jmse9020156.
[42] Chen E, Bao H, Dinh N. Evaluating the reliability of machine-learning-based predictions used in nuclear power plant instrumentation and control systems. Reliab Eng Syst Saf 2024;250:110266. https://doi.org/10.1016/j.ress.2024.110266.
[43] Palar PS, Zuhal LR, Shimoyama K. Enhancing the explainability of regression-based polynomial chaos expansion by shapley additive explanations. Reliab Eng Syst Saf 2023;232:109045. https://doi.org/10.1016/j.ress.2022.109045.
[44] Goda T. A simple algorithm for global sensitivity analysis with shapley effects. Reliab Eng Syst Saf 2021;213:107702. https://doi.org/10.1016/j.ress.2021.107702.
[45] Liu KZ, Yu Q, Yuan ZT, Yang ZS, Shu YQ. A systematic analysis for maritime accidents causation in Chinese coastal waters using machine learning approaches. Ocean Coast Manag 2021;213:105859. https://doi.org/10.1016/j.ocecoaman.2021.105859.
[46] Wang HX, Liu ZJ, Wang XJ, Graham T, Wang J. An analysis of factors affecting the severity of marine accidents. Reliab Eng Syst Saf 2021;210:107513. https://doi.org/10.1016/j.ress.2021.107513.
[47] Yan R, Wang SA, Peng CS. An artificial intelligence model considering data imbalance for ship selection in port state control based on detention probabilities. J Comput Sci 2021;48:101257. https://doi.org/10.1016/j.jocs.2020.101257.
[48] Bagci Das D. Real-time adaptable fault analysis of rotating machines based on marine predator algorithm optimised LightGBM approach. Nondestruct Test Eval 2025;40(3):831–66. https://doi.org/10.1080/10589759.2024.2328750.
[49] Li T, Wang X, Zhang Z, Feng Y. A novel feature engineering method for severity prediction of marine accidents. J Mar Eng Technol 2025:1–16. https://doi.org/10.1080/20464177.2025.2469336.
[50] Bei H, Yang F, Wang W, Yang T, Murcio R. Maritime safety through multi-source data fusion: an AdaBoost-based approach for predictive ship detention by port state control. Marit Policy Manag 2024:1–20. https://doi.org/10.1080/03088839.2024.2438901.
[51] Chen T, Wang H, Cai Y, Liang M, Meng Q. Exploring key factors for long-term vessel incident risk prediction. Reliab Eng Syst Saf 2025;253:110565. https://doi.org/10.1016/j.ress.2024.110565.
[52] Chen T, Wong YD, Shi X, Wang X. Optimized structure learning of Bayesian network for investigating causation of vehicles' on-road crashes. Reliab Eng Syst Saf 2022;224:108527. https://doi.org/10.1016/j.ress.2022.108527.
[53] Mostafaei S, Ahmadi A, Shahrabi J. USWAVG-BS: under-sampled weighted AVeraGed BorderlineSMOTE to handle data intrinsic difficulties. Expert Syst Appl 2023;227:120379. https://doi.org/10.1016/j.eswa.2023.120379.
[54] Lin CJ, Leony F. Evidence-based adaptive oversampling algorithm for imbalanced classification. Knowl Inf Syst 2024;66(3):2209–33. https://doi.org/10.1007/s10115-023-01985-5.
[55] Alhudhaif A, Almaslukh B, Aseeri AO, Guler O, Polat K. A novel nonlinear automated multi-class skin lesion detection system using soft-attention based convolutional neural networks. Chaos Solit Fractals 2023;170:113409. https://doi.org/10.1016/j.chaos.2023.113409.
[56] Wang Z, Wu CH, Zheng KF, Niu XX, Wang XJ. SMOTETomek-based resampling for personality recognition. IEEE Access 2019;7:129678–89. https://doi.org/10.1109/ACCESS.2019.2940061.
[57] Guo R, Guo J, Zhang LC, Qu XX, Dai SF, Peng RC, Chong VFH, Xian JF. CT-based radiomics features in the prediction of thyroid cartilage invasion from laryngeal and hypopharyngeal squamous cell carcinoma. Cancer Imaging 2020;20(1):81. https://doi.org/10.1186/s40644-020-00359-2.
[58] Wichitaksorn N, Kang Y, Zhang F. Random feature selection using random subspace logistic regression. Expert Syst Appl 2023;217:119535. https://doi.org/10.1016/j.eswa.2023.119535.
[59] Panhalkar AR, Doye DD. A novel approach to build accurate and diverse decision tree forest. Evol Intell 2022;15(1):439–53. https://doi.org/10.1007/s12065-020-00519-0.
[60] Rhodes JS, Cutler A, Moon KR. Geometry- and accuracy-preserving random forest proximities. IEEE Trans Pattern Anal Mach Intell 2023;45(9):10947–59. https://doi.org/10.1109/TPAMI.2023.3263774.
[61] Alsariera YA, Adeyemo VE, Balogun AO, Alazzawi AK. AI meta-learners and extra-trees algorithm for the detection of phishing websites. IEEE Access 2020;8: 142532–42. https://doi.org/10.1109/ACCESS.2020.3013699.
[62] Liu R. Research on Fault Diagnosis Based on Artificial Neural Network. Singapore: Springer Singapore; 2020. p. 73–9.
[63] Liang JC, Bu YD, Tan KF, Pan JC, Yi ZP, Kong XM, Fan Z. Estimation of stellar atmospheric parameters with light gradient boosting machine algorithm and principal component analysis. Astron J 2022;163(4):153. https://doi.org/10.3847/1538-3881/ac4d97.
[64] Yan ZZ, Chen H, Dong XH, Zhou KW, Xu ZG. Research on prediction of multi-class theft crimes by an optimized decomposition and fusion method based on XGBoost. Expert Syst Appl 2022;207:117943. https://doi.org/10.1016/j.eswa.2022.117943.
[65] Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data 2020;7(1):94. https://doi.org/10.1186/s40537-020-00369-8.
[66] Kosugi Y, Mizuno K, Santos C, Sato S, Hosea N, Zientek M. Direct comparison of the prediction of the unbound brain-to-plasma partitioning utilizing machine learning approach and mechanistic neuropharmacokinetic model. AAPS J 2021;23(4):72. https://doi.org/10.1208/s12248-021-00604-x.
[67] Mahoto NA, Shaikh A, Sulaiman A, Reshan MSA, Rajab A, Rajab K. A machine learning based data modeling for medical diagnosis. Biomed Signal Process Control 2023;81:104481. https://doi.org/10.1016/j.bspc.2022.104481.
[68] Rajeshwar TR, Demerdash ONA, Smith JC. TCR-H: explainable machine learning prediction of T-cell receptor epitope binding on unseen datasets. Front Immunol 2024;15:1426173. https://doi.org/10.3389/fimmu.2024.1426173.
[69] Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform 2014;6: 10. https://doi.org/10.1186/1758-2946-6-10.
[70] Mutz J, Iniesta R, Lewis CM. Metabolomic age (MileAge) predicts health and life span: a comparison of multiple machine learning algorithms. Sci Adv 2024;10(51): 3743. https://doi.org/10.1126/sciadv.adp3743.
[71] Rawson A, Brito M, Sabeur Z, Tran-Thanh L. A machine learning approach for monitoring ship safety in extreme weather events. Saf Sci 2021;141:105336. https://doi.org/10.1016/j.ssci.2021.105336.
[72] Upadhyaya DP, Tarabichi Y, Prantzalos K, Ayub S, Kaelber DC, Sahoo SS. Machine learning interpretability methods to characterize the importance of hematologic biomarkers in prognosticating patients with suspected infection. Comput Biol Med

2024;183:109251. https://doi.org/10.1016/j.compbiomed.2024.109251. -109251.
[73] Zhou KW, Xing WB, Wang JB, Li HH, Yang ZL. A data-driven risk model for maritime casualty analysis: a global perspective. Reliab Eng Syst Saf 2024;244: 109925. https://doi.org/10.1016/j.ress.2023.109925.
[74] Li GL, Wu YD, Bai YL, Zhang WH. ReMAHA-CatBoost: addressing imbalanced data in traffic accident prediction tasks. Appl Sci 2023;13(24):13123. https://doi.org/10.3390/app132413123.
[75] Ge XT, Fang CR, Bai TT, Liu J, Zhao ZH. An empirical study of class rebalancing methods for actionable warning identification. IEEE Trans Reliab 2023;72(4): 1648–62. https://doi.org/10.1109/TR.2023.3234982.
[76] Chen HY, Shen QG, Skibniewski MJ, Cao Y, Liu Y. Dynamic prediction and optimization of tunneling parameters with high reliability based on a hybrid intelligent algorithm. Inf Fusion 2024;114:102705. https://doi.org/10.1016/j.inffus.2024.102705.
[77] Yue H. Investigating the influence of streetscape environmental characteristics on pedestrian crashes at intersections using street view images and explainable machine learning. Accid Anal Prev 2024;205:107693. https://doi.org/10.1016/j.aap.2024.107693.
[78] Weng JX, Yang D, Qian T, Huang Z. Combining zero-inflated negative binomial regression with MLRT techniques: an approach to evaluating shipping accident casualties. Ocean Eng 2018;166:135–44. https://doi.org/10.1016/j.oceaneng.2018.08.011.
[79] Cakir E, Sevgili C, Fiskin R. An analysis of severity of oil spill caused by vessel accidents. Transp Res D-Transp Environ 2021;90:102662. https://doi.org/10.1016/j.trd.2020.102662.
[80] Goksu S, Arslan O. A quantitative dynamic risk assessment for ship operation using the fuzzy FMEA: the case of ship berthing/unberthing operation. Ocean Eng 2023; 287:115548. https://doi.org/10.1016/j.oceaneng.2023.115548.
[81] Swateelagna S, Singh M, Rahul MR. Explainable machine learning based approach for the design of new refractory high entropy alloys. Intermetallics 2024;167: 108198. https://doi.org/10.1016/j.intermet.2024.108198 (Barking).
[82] Kim J, Yoo SL. Coastal traffic safety index based on marine accident and traffic records. J Mar Sci Technol 2023;31(3):283. https://doi.org/10.51400/2709-6998.2704.
[83] Elidolu G, Teixeira AP, Arslanoglu Y. A risk assessment of inhibited cargo operations in maritime transportation: a case of handling styrene monomer. Ocean Eng 2024;312:119049. https://doi.org/10.1016/j.oceaneng.2024.119049.