Feng, Y., Wang, H., Xia, G., Cao, W., Li, T., Wang, X., Liu, Z., 2025. A machine learning-based data-driven method for risk analysis of marine accidents. *Journal of Marine Engineering & Technology*, 24 (2), 147-158.
https://doi.org/10.1080/20464177.2024.2368914

A machine learning-based data-driven method for risk analysis of marine accidents

Yinwei Feng [a, b], Huanxin Wang [a, b, c, d *], Guoqing Xia [a, d], Wenjie Cao [a, d], Tianyi Li [a, d], Xinjian Wang [a, b, c, d], Zhengjiang Liu [a, d]

[a] *Navigation College, Dalian Maritime University, Dalian 116026, P.R. China*
[b] *Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool L3 3AF, UK*
[c] *Key Laboratory of International Shipping Development and Property Digitization of Hainan Free Trade Port, Hainan Vocational University of Science and Technology, Haikou 570100, PR China*
[d] *Key Laboratory of Navigation Safety Guarantee of Liaoning Province, Dalian 116026, P.R. China*

Abstract:

In view of the frequent occurrence of marine accidents and the complex interaction of various risk-influencing factors (RIFs), a data-driven method to risk analysis that combines association rule mining (ARM) and complex network (CN) analysis is proposed in this study. The efficient FP-Growth algorithm is applied to facilitate ARM to examine risk patterns that frequently occur in marine accidents. Subsequently, CN theory is employed to scrutinize the multifaceted role of various RIFs and their interactions in the complex marine accident system, which involves the basic characteristics of the network, the identification of key RIFs through the application of the weighted LeaderRank (WLR) algorithm, and a robustness analysis. The results of the study indicate that compared with random networks, marine accident networks exhibit a higher level of complexity, which brings challenges to safety prevention and control. Inadequate regulation, violations, and deficiencies in safety management systems are identified as key RIFs, stressing the urgency of improving supervision, strengthening law enforcement and strengthening the safety management system. This study may facilitate maritime safety management of maritime traffic and the development of risk analysis methods.

Keywords: Maritime safety, Marine accident, Machine learning, Association rule mining, Complex network

* *Corresponding author:* *wanghxdmu@dlmu.edu.cn* *(H. Wang)*

1. Introduction

Seaborne trade plays a pivotal role in global economic development by providing a cost-effective and environmentally friendly means of efficiently transporting passengers and cargo between locations (Cao *et al.*, 2023a; Christensen *et al.*, 2022; Fang *et al.*, 2024; Li *et al.*, 2023a). However, maritime shipping stands out as one of the most perilous industries globally (IMO, 2023), characterized by a consistently high incidence of marine accidents throughout the year. The annual overview of marine casualties and incidents in 2022 reported an average of 2,647 casualties and incidents per year over the period 2014-2021 (EMSA, 2022). A similar maritime safety landscape is evident in the Allianz Safety & Shipping Review 2023 (AGCS, 2023), which documented 3,032 shipping incidents in 2022. Despite concerted efforts, the persistent challenge in the shipping community lies in the control of risks and the assurance of shipping safety, considering the high frequency of incidents and the consequential serious casualties and economic losses caused by marine accidents (Cao *et al.*, 2023a; Shafiee and Animah, 2022; Wang *et al.*, 2022a).

To address the challenges inherent in maritime safety, numerous studies have been undertaken to analyze the causes, development, evolution laws, and strategies for the prevention and control of marine accidents and incidents (Wang *et al.*, 2023b; Zhang *et al.*, 2021; Zhao *et al.*, 2023). Several investigations have systematically identified and analyzed the key risk-influencing factors (RIFs) of marine accidents across diverse scenarios by examining historical data pertaining to marine accidents (Cao *et al.*, 2023b; Wang *et al.*, 2022c). Additionally, certain studies (Cao *et al.*, 2023a; Li *et al.*, 2023b) posit that maritime safety is intricately linked not only to specific RIFs characterized by high uncertainty, such as the human element, ship conditions, and environmental conditions, but also to their intricate interplay (Kimera and Nangolo, 2022; Valcalda *et al.*, 2023). Furthermore, it is noted that the impact of these RIFs on maritime safety is dynamic, and evolves over time. Any latent hazard stemming from these RIFs have the potential to propagate within the system and culminate in accidents ultimately. Therefore, this study employs a methodology integrating association rules mining (ARM) and complex network (CN) to analyze marine accident RIFs, aiming to deeply study the dynamic interaction among marine accident RIFs, grasp the mechanisms and patterns of risk transmission, and identify key RIFs in the system.

The subsequent sections of this study are structured as follows. Section 2 delineates the research gaps by analyzing prior studies on the risk analysis of marine accidents. Section 3 provides an overview of the data sources, the fundamental concepts of each method, and the research framework. In Section 4, the RIFs of marine accidents are extracted through ARM, followed by the construction and analysis of a marine accident association network using CN analysis. In Section 5, specific recommendations for risk mitigation strategies are provided based on the identified interactions of RIFs in marine accidents. The conclusions of this study are presented in Section 6.

2. Literature review

The occurrence of marine accidents is the result of a combination of various RIFs (Cao *et al.*, 2023b). Therefore, the identification and quantitative assessment of these RIFs are key focal points in the research on marine accident risk analysis (Zhang *et al.*, 2022). After several years of development, existing research has systematically categorized the RIFs contributing to marine accidents into five main categories: human factors, vessel factors, environmental factors, management factors (also referred to as organizational factors), and accident information (such as time and type of accident) (Fan and Yang, 2024; Tian *et al.*, 2020). Researchers have employed a variety of methods to further explore the impact of these RIFs on marine accidents within these aforementioned categories (Cao *et al.*, 2023a; Li *et al.*, 2023b; Wang *et al.*, 2021; Xia *et al.*, 2023). Weng *et al.* (2020) introduced a ship dynamic domain model to evaluate the probability of ship collisions. This model considered RIFs such as the ship's angle of approach, ship type, relative speed, and day and night conditions. The findings indicated that narrower channel widths, adverse weather conditions, and darkness significantly impact the likelihood of ship collisions. Hwang and Youn (2022) conducted a data-driven text analysis on accident investigation reports to establish connections between collision accidents and other incidents resulting from collisions. Their analysis sought to uncover RIFs contributing to the evolution of collision accidents into other types of accidents, with the goal of preventing further escalation of accident risks. Ugurlu and Cicek (2022) delved into the RIFs influencing ship accidents by analyzing numerous accident reports. Their application of fault tree (FT) analysis identified human error as the predominant factor in ship accidents, leading to corresponding recommendations focused on the crew's role in accident prevention. Wang *et al.* (2022b) undertook a comprehensive review spanning 26 years of tanker accidents. Using the binomial regression model and kernel density estimation, the primary RIFs contributing to accidents and accident intensity points were identified respectively. Cakir *et al.* (2021) employed decision trees (DT) and data-driven Bayesian networks (BNs) to examine RIFs associated with the severity of oil spill accidents. Notably, their study revealed that non-US flagged vessels were more prone to causing serious oil spills compared to US flagged vessels. Additionally, sinking and grounding emerged as the most perilous types of accidents leading to severe oil spills.

The present corpus of research, rooted in classical accident analysis theories, is effective in enhancing the understanding of the mechanisms underlying marine accidents. However, when applied to a wide range of analyses, it becomes apparent that these approaches are not only time-consuming and expensive but also susceptible to incomplete and biased causal factors stemming from individual experiences and knowledge. For instance, BNs, although proficient in analyzing coupling relationships among RIFs, require high data quality and depend on expert knowledge to determine certain priori probabilities. Given these constraints, researchers in safety and risk management are increasingly turning to systemic accident association models for their investigative pursuits. It has been proven that the utilization of association rules in the realm of maritime safety helps to identify latent safety hazards and correlations between accidents (Huang *et al.*, 2023;

Lan *et al.*, 2023; Zhou *et al.*, 2019b). For example, Lan *et al.* (2023) utilized ARM to investigate significant patterns contributing to the total loss of marine accidents, and employed the Éclat algorithm to extract association rules related to accident types and severity. The research findings indicated that vessel age exceeding 20 years was a crucial indicator contributing to casualties. Hull/machinery damage and sinking emerged as the primary accident types leading to total loss of ships. Huang and Hu (2019) employed the Apriori algorithm to ARM on marine accident data, thereby generating strong association rules among RIFs of marine accidents. The results revealed significant associations between marine accidents and crew errors, as well as those between marine accidents and environmental factors. Yu *et al.* (2021) utilized ARM to examine the relationships and distinctions among RIFs influencing accidents in various risk regions. The findings indicated that small ship size, general cargo type, and spring or summer seasons exhibited a significant association with collisions in high-risk regions. Huang *et al.* (2023) employed weighted-ARM (WARM) to analyze the causal characteristics of marine accidents in the Mediterranean Sea and the Black Sea, revealing a robust correlation between accidents and ship age, gross tonnage, and ship type.

Although those existing studies have conducted systematic analyses of the mechanisms of marine accidents, there are still numerous unresolved issues, particularly the complex dependencies and interactions among accident-related RIFs that remain inadequately interpretation. Moreover, the current identification of RIFs is not comprehensive, which hinders an in-depth analysis of accident mechanisms. Recognizing these research gaps, this study proposes a hybrid model integrating ARM and CN analyses, aiming to systematically identify the key RIFs associated with various types of marine accidents and their interaction mechanisms. The contributions of this study are as follows:

(1) A date-driven methodological framework integrating ARM technology and CN was proposed to conduct a deep exploration of the intrinsic associations, including co-occurrence, correlation, or causation, among the concentration of RIFs in marine accidents.

(2) The associations were mapped into a CN, and the nodes in the CN were ranked based on their importance using the advanced Weighted LeaderRank (WLR) algorithm, facilitating a comprehensive understanding of their interrelationships, as well as reflecting the criticality of various RIFs.

(3) The methodological framework of this study provides a new perspective for maritime safety management, enabling decision-makers to make wiser decisions in complex navigational environments, thereby effectively reducing accident rates and ensuring the safety and reliability of maritime transportation.

3. Methodology

This study innovatively integrates ARM technology and CN theory to formulate and scrutinize the association network of marine accidents. The procedural depiction of utilizing CN, based on ARM for risk analysis, is illustrated in Figure 1. Firstly, significant accident RIFs are extracted from extensive accident reports and their hierarchical classification is consistent with the risk management model. Subsequently, the ARM method is applied to reveal robust association rules among RIFs, providing the foundation for constructing the marine accident association network. On this basis, the CN theory is used to identify the contributing RIFs and their intricate interaction mechanisms, and the core factors and association links are determined. Finally, the key RIFs are validated through the network robustness analysis.

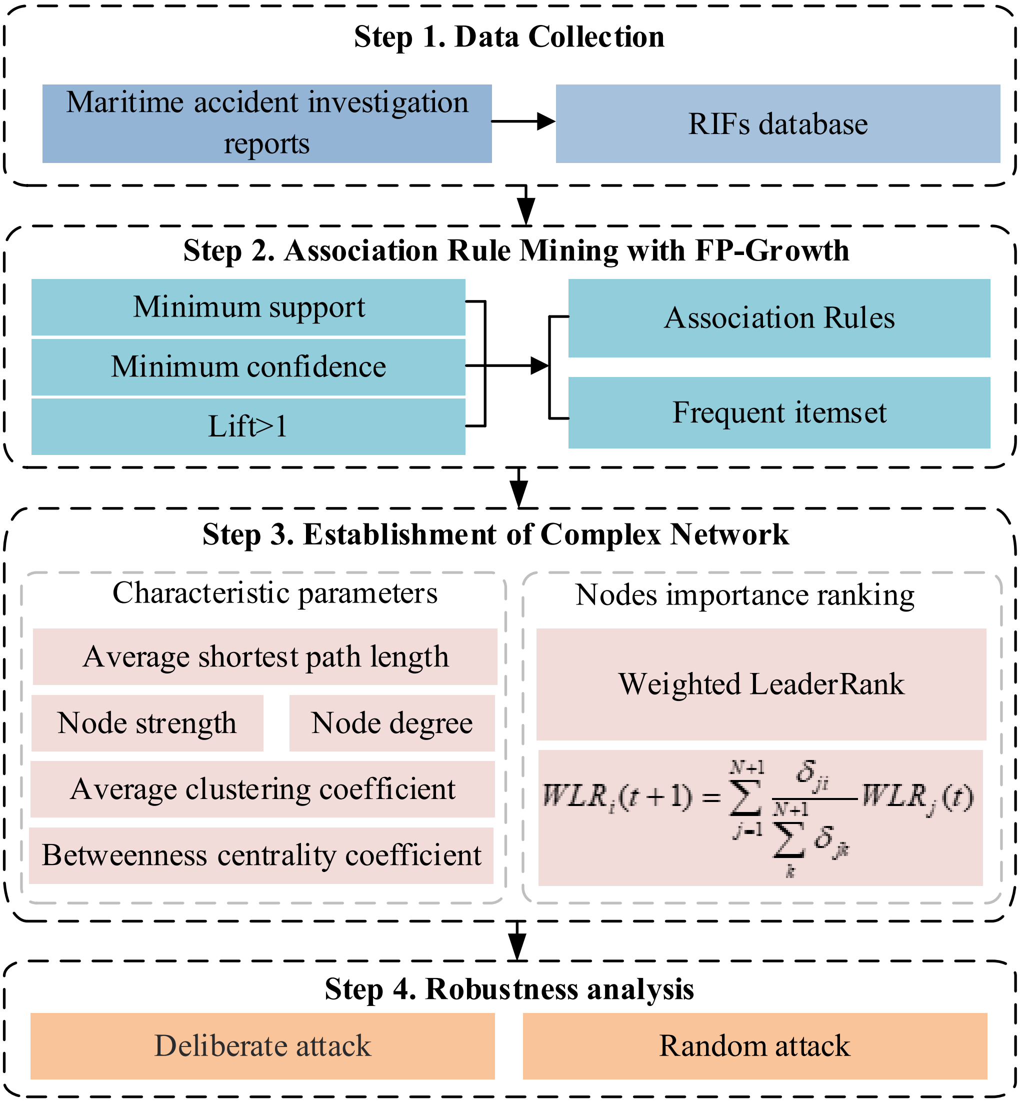

Figure 1 Processes for marine accident risk analysis using ARM and CN.

3.1 Data collection and preprocessing

Marine accident investigation reports provide detailed information about each accident investigated, including when and where it occurred, the people and the vessel involved. This information helps researchers to understand the full picture of accidents and identify the direct and deep causes of accidents, such as human error, equipment failure, mismanagement and so on. Therefore, these investigation reports are usually considered as the most detailed and reliable data source for accident analysis.

In this study, 1207 marine accident investigation reports were collected, which were published by different agencies, such as Marine Accident Investigation Branch (MAIB) and National Transport Safety Board (NTSB). The sources of the investigation reports are shown in Figure 2. The accidents included in these reports were occurred in the period from 2000 to 2019. Although the presentation forms of the investigation reports of different investigation agencies are different, the main elements contained in each investigation report are basically the same. Each record contains the information regarding accident narrative, crew characteristics, ship condition, environmental condition, and management characteristics. More detailed information with regard to these RIFs is listed below.

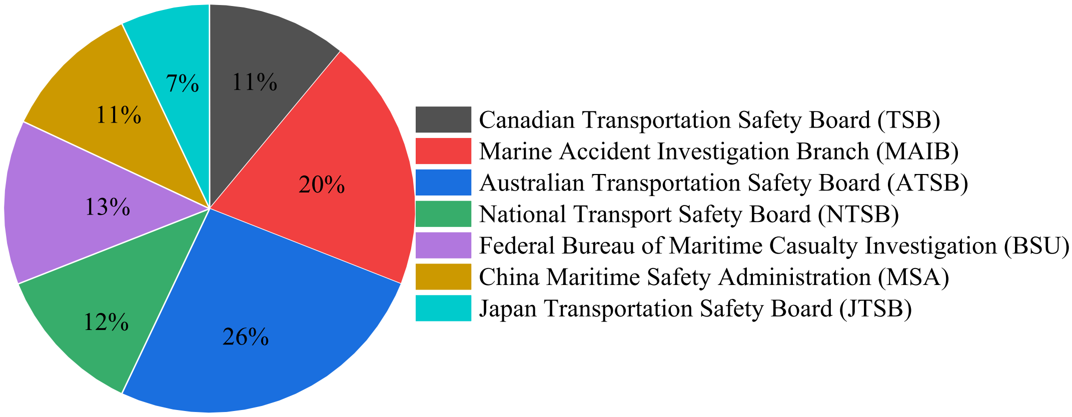

Figure 2 The sources of the investigation reports.

(a) Accident characteristics: accident type and accident time are taken into account as accident characteristics. Accident types are divided into seven categories in this study, which include collision, grounding, fire/explosion, contact, sinking, equipment failure and others (e.g., hull damage/machinery failure) etc.

(b) Human characteristics: detailed information with regard to experience, education, human erroneous action and violation etc.

(c) Ship characteristics: this type of information includes gross tonnage, engine power and ship type (cargo/container ship; fishing boat; Oil tanker; other ship types). It should be pointed out that towing vessels, engineering ships and sand dredgers belong to the category of other ship types. The factors under the umbrella of ship conditions include ship age, ship manning, the ship's certificates and seafarers' certificates etc.

(d) Environmental characteristics: this type of information includes detailed information with regard to the visibility (poor visibility condition; very poor visibility condition), winds (strong winds; very strong winds; Storm), and traffic density etc.

(e) Management characteristics: this type of information encompasses the guidelines, regulations, and regulatory initiatives established by the governing body, in addition to the safety management system, crew training, and corporate safety culture within the maritime company etc.

To ensure the accuracy and efficacy of the analysis results, the study identifies and extracts five categories of RIFs: accident characteristics, human elements, ship types, ship conditions, and environmental conditions. Subsequently, the collected data undergoes a comprehensive preprocessing phase, encompassing activities such as data cleaning, integration, and transformation. The original dataset, being unstructured, presents a diverse array of information, leading to challenges like data gaps, noise, and inconsistencies. To mitigate the impact of noise on the analysis outcomes, data with missing values is excluded, and the existing data is integrated and transformed into a structured format (processing details are detailed in previous studies (Cao *et al.*, 2023a; Wang *et al.*, 2021)). For detailed descriptions of these RIFs, please consult Table A in the Appendix.

3.2 Association rule mining

ARM stands out as a frequently employed technique within unsupervised learning algorithms (Nguyen *et al.*, 2019). Primarily utilized to unveil correlations among diverse items within extensive datasets, ARM aims to extract frequent patterns, associations, co-occurrences, or causal relationships among a multitude of variables (Zhang *et al.*, 2018). The identification of rules hinges on satisfying predetermined criteria, specifically the minimum support threshold and minimum confidence degree, leading to the discovery of relevant association rules (Solanki and Patel, 2015).

The association rules concept is as follows: Let $A = \{a_1, a_2, ..., a_n\}$ be itemset and $D = \{d_1, d_2, ..., d_n\}$ be the n transaction database, where $d_i \subseteq A$. Ar contains n items, so that Ar is called a frequent n-itemset. A set of association rules output from ARM can be represented as $X \Rightarrow Y$, where $X \subset I, Y \subset I, X \cap Y = \varnothing$. X and Y are called antecedent and consequent transactions, respectively.

As illustrated in Table 1, RIFs of marine accidents (including attributes such as ship type, accident time, and navigational environment) collectively form an itemset. "Collision", "Inland waters" and "Strong wind" each serve as an item and ["Collision"; "Inland waters"; "Strong wind"] constitutes a 3-itemset. ["Time: 0000-0400"; "Collision"; "Inland waters"; "Strong wind"; "Inadequate regulation"; "Container ship"] form a transaction, with a corresponding ID value equal to 1.

Table 1 Example table of marine accident information dataset.

ID	Transaction
1	Time: 0000-0400; Collision; Inland waters; Strong wind; Inadequate regulation; Container ship
2	Time: 0400-0800; Contact; Coastal waters; Very strong wind; Inadequate supervision, Chemical tanker
3	Time: 0800-1200; Collision; Port; Very strong wind; Inadequate supervision; Fishing vessel
…	……
n	Time: 2000-2400; Fire/Explosion; Open Sea; Strom; Inadequate ship training; Yacht and sailing vessel

The gauging of the robustness or significance of association rules commonly relies on metrics such as support and confidence (Nguyen *et al.*, 2019). An association rule is deemed potent when it satisfies predefined thresholds for minimum support and minimum confidence. However, the practical meaningfulness of association rules with high potency is not assured, prompting the use of lift as a measure of their value. A valuable association rule is identified when the lift exceeds 1 (Lee and Yu, 2023). The computation of support, confidence, and lift for association rules is detailed in Equations (1) to (3):

$$Supp(X) = \frac{M_X}{M} \tag{1}$$

$$Con(X \Rightarrow Y) = \frac{Supp(XY)}{Supp(X)} \tag{2}$$

$$Lift(X \Rightarrow Y) = \frac{Con(X \Rightarrow Y)}{Supp(Y)} \tag{3}$$

where $Supp(X)$ denotes the support of transaction X; $Con(X \Rightarrow Y)$ denotes the confidence level of $X \Rightarrow Y$; $lift(X \Rightarrow Y)$ denotes the lift level of $X \Rightarrow Y$; M_X indicates the number of occurrences of X transactions in the itemset; M indicates the number of transactions in the itemset.

ARM process comprises two primary phases. In the initial phase, the objective is to identify all frequent itemsets within the original database. Subsequently, in the second phase, association rules are generated based on the identified frequent itemsets. It is worth noting that the computational complexity of ARM experiences exponential growth as the number of association rules increases (Li *et al.*, 2023b; Nguyen *et al.*, 2019). To tackle this challenge, rather than opting for Apriori, a widely employed algorithm in numerous studies, this research has selected FP-Growth, an algorithm renowned for its operational efficiency (Narvekar and Syed, 2015). Figure 3 illustrates the specific flow of the FP-Growth algorithm (Çakır *et al.*, 2021).

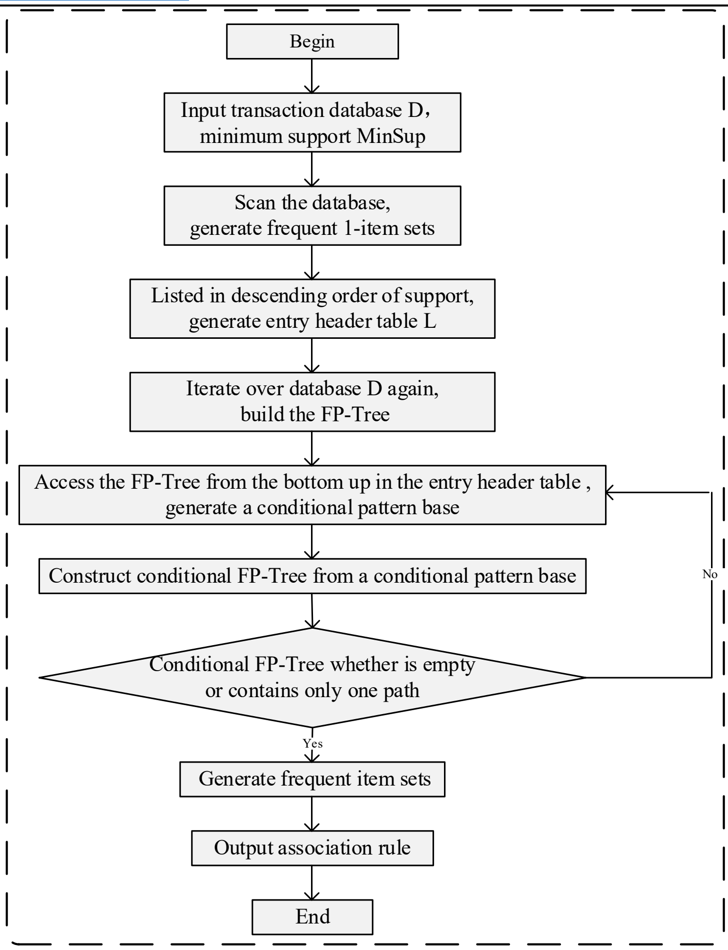

Figure 3 Flow chart of FP-Growth algorithm.

3.3 Weighted directed Complex network

CNs represent a distinctive structural model, abstracting elements within a complex system as nodes and interconnections between elements as edges. CN has found extensive application in scrutinizing the interplay of RIFs in various domains, including railroad and road traffic accidents (Fu *et al.*, 2022; Shi *et al.*, 2024). In an effort to elucidate the interaction relationships among the RIFs of marine accidents, this study employs CN methodology for the analysis of RIFs within this context. The extraction of association rules among the RIFs contributes to the construction of a weighted, directed-CN. The association rules among the RIFs are transcribed onto a CN, as illustrated in Figure 4. The antecedent represents the originating node, the consequent designates the destination node, and the confidence parameter signifies the weight of the edges connecting these nodes.

Let the interaction network of RIFs be $DiG = (N, DiE)$, where $N = \{n_1, n_2, ..., n_p\}$ denotes the set of nodes of the network and $DiE = \{die_1, die_2, ..., die_q\}$ denotes the set of edges of the network. The mathematical expression of the network can be described as the adjacency matrix G, with Equations (4) to (6).

$$G = \begin{array}{c} \\ n_1 \\ n_2 \\ \vdots \\ n_p \end{array} \begin{array}{c} \begin{array}{cccc} n_1 & n_2 & \cdots & n_p \end{array} \\ \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1p} \\ g_{21} & g_{22} & \cdots & g_{2p} \\ \vdots & \vdots & & \vdots \\ g_{p1} & g_{p2} & \cdots & g_{pp} \end{bmatrix} \end{array} \tag{4}$$

$$g_{ij} = Con(n_i \Rightarrow n_j) \times e_{ij} \tag{5}$$

$$e_{ij} = \begin{cases} 1, \ when\ the\ i\ event\ triggers\ the\ j\ event \\ 0, \ else \end{cases} \tag{6}$$

where g_{ij} denotes the connectivity from node n_i to node n_j; $Con(n_i \Rightarrow n_j)$ denotes the weight of the edges from node n_i to node n_j in the network; e_{ij} denotes the connectivity from node n_i to node n_j.

Antecedent	Consequent	Confidence
A	B	1.00
C	B	0.98
C	E	0.97
D	A	0.92
E	D	0.88

	A	B	C	D	E
A		1.00	0	0	0
B	0		0	0	0
C	0	0.98		0	0.97
D	0.92	0	0		0
E	0	0	0	0.88	

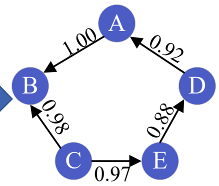

Figure 4 Mapping process from association rules to a weighted directed-CN.

3.3.1 Characteristic parameters

The defining parameters of networks encompass the statistical distribution of microscopic variables and the mean of macroscopic statistics within a designated network (Zhou *et al.*, 2019b). This forms the foundation for scrutinizing the topological attributes of complex networks. Parameters characterizing weighted directed networks predominantly encompass node degree, node strength, betweenness centrality, clustering coefficient, and other relevant metrics (Feng *et al.*, 2021; Fu *et al.*, 2022).

(1) Node degree

The node degree is the total number of edges of the node. Degree of a directed network is divided into node total-degree, node out-degree, and node in-degree. Total-degree of node n_i represents the number of edges connecting node n_i. Out-degree of node n_i represents the number of edges from node n_i to other nodes. In-degree of node n_i represents the number of edges from other nodes to node n_i. The calculation formulas are as in Equations (7) to (9):

$$k_i^{in} = \sum_{j=1}^{p} e_{ji} \tag{7}$$

$$k_i^{out} = \sum_{j} e_{ij} \tag{8}$$

$$k_i^{total} = k_i^{in} + k_i^{out} \tag{9}$$

where k_i^{in} denotes the in-degree of node n_i; k_i^{out} denotes the outgoing degree of node n_i; and k_i^{total} denotes the total-degree of node n_i.

(2) Node strength

The node strength is a unique metric for weighted networks, which refers to the sum of the weights of the edges connecting the nodes. Node strength of a directed network is divided into node total-strength, node out-strength, and node in-strength. Total-strength of node n_i represents the sum of the weights of the edges of node n_i. Out-strength of node n_i represents the sum of weights from node n_i to other nodes. In-strength of node n_i represents the sum of weights from other nodes to node n_i. The calculation formulas are as in Equations (10) to (12):

$$S_i^{in} = \sum_{j,j\neq i} g_{ji} \tag{10}$$

$$S_i^{out} = \sum_{j,j\neq i} g_{ij} \tag{11}$$

$$S_i^{total} = S_i^{in} + S_i^{out} \tag{12}$$

where S_i^{in} denotes the in-strength of node n_i; S_i^{out} denotes the out-strength of node n_i; and S_i^{total} denotes the total-strength of node n_i.

(3) Node betweenness centrality coefficient

The metric known as node betweenness centrality coefficient quantifies the number of shortest paths passing through a given node within the network. This metric provides insights into the centrality and transmission capabilities of the node, and its calculation is expressed as follows:

$$BC_v = \frac{1}{(p-1)(p-2)} \sum_{\substack{\forall i,j\in N \\ i\neq v\neq j}} \left(\frac{\sigma_{ij}(v)}{\sigma_{ij}}\right) \tag{13}$$

where $\sigma_{ij}(v)$ denotes the number of shortest paths from node n_i to node n_j through node n_v; σ_{ij} denotes the number of shortest paths from node n_i to node n_j; p denotes the number of nodes in the network; BC_v denotes the betweenness centrality of node n_v.

(4) Average shortest path length

The network's average shortest path length (ASPL) denotes the mean shortest path length between any two nodes in the network. The calculations for ASPL are expressed as follows:

$$ASPL = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} l_{ij}}{p(p-1)} \tag{14}$$

where l_{ij} denotes the shortest path length from node n_i to n_j.

(5) Average clustering coefficient (ACC)

The mean clustering coefficient of the network reflects the level of node aggregation within the network and is computed using the following formula:

$$ACC = \frac{1}{p}\sum_{i}^{p}\frac{2E_i}{k_i^{total}(k_i^{total}-1)} \tag{15}$$

where E_i denotes the number of edges of node n_i's neighbours' nodes.

3.3.2 Nodes importance ranking

WLR is an algorithm that uses feature vector centrality to rank the importance of nodes in CNs (Li *et al.*, 2014). The main idea of the algorithm is to add a common node Ground to a known CN $DiG=(N,DiE)$, thus obtaining a directed network $DiG'=(N+1,DiE+2N)$ with $N+1$ nodes. The reciprocal links connecting the "Ground" with other nodes establish a strongly connected graph, thereby preventing the existence of isolated nodes in a CN and ensuring the convergence of the algorithm. Furthermore, the algorithm enhances the iterative formula of the initial LeaderRank by taking into account network characteristics (Li *et al.*, 2014). Specifically, nodes with higher in-degrees accrue greater scores during the iterative process. In addition, the algorithm improves the iterative formula of the original LeaderRank based on the network characteristics. WLR algorithm network structure is shown in Figure 5. The calculation formulas are as in Equations (16):

$$WLR_i(t+1) = \sum_{j=1}^{N+1}\frac{\delta_{ji}}{\sum_{k}^{N+1}\delta_{jk}}WLR_j(t) \tag{16}$$

where t denotes the number of iterations; δ_{ij} denotes the weights. If $e_{ij}=0$, then $\delta_{ij}=0$; if n_j is Ground, then $\delta_{gi}=(k_i^{in})^{\alpha}$, where α is a free parameter; for all other cases $\delta_{ij}=1$.

	A	B	C	D	E
A		1.00	0	0	0
B	0		0	0	0
C	0	0.98		0	0.97
D	0.92	0	0		0
E	0	0	0	0.88	

	A	B	C	D	E	G
A		1.00	0	0	0	1
B	0		0	0	0	1
C	0	1.00		0	1.00	1
D	1.00	0	0		0	1
E	0	0	0	1.00		1
G	1	2	0	1	1	

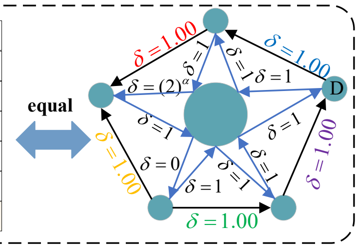

Figure 5 WLR algorithm network structure.

3.4 Robustness analysis

The robustness of the CN refers to its capacity to stay interconnected even when specific nodes or edges experience failure. Robustness serves as an indicator of the network's stability and continuity to some degree (Wang *et al.*, 2023a). The disruption of the interactive network structure involving RIFs in marine traffic accidents and the severance of this network are pivotal in impeding accident progression and facilitating accident prevention. In this study, CN stability (i.e., a combination of reachability and global efficiency) is used as a metric to evaluate network robustness. The calculation formulas are as in Equations (17) to (20):

$$Eff = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{1}{l_{ij}}}{p(p-1)} \tag{17}$$

$$RP = \begin{bmatrix} RP_{11} & \cdots & RP_{1p} \\ \vdots & & \vdots \\ RP_{p1} & \cdots & RP_{pp} \end{bmatrix} \tag{18}$$

$$R = (ones)_{1\times p} \cdot RP \cdot (ones)_{p\times 1} \tag{19}$$

$$NS = ((\frac{Eff_{attack}}{Eff_o}) \times (\frac{R_{attack}}{R_o}))^{\frac{1}{2}} \tag{20}$$

where Eff denotes the global efficiency of the network; RP denotes the reachability matrix of the network; RP_{ij} denotes the connectivity of node n_i to n_j. If there exists at least one shortest path from node n_i to n_j, then RP_{ij} equals 1; otherwise RP_{ij} equals 0; R denotes the network reachability; $(one)_{1\times p}$ and $(one)_{p\times 1}$ denote the all-one matrix of $1\times p$ and the matrix of $p\times 1$, respectively; Eff_{attack} and R_{attack} denote the global efficiency and reachability of the network after being attacked, respectively; Eff_o and R_o denote the global efficiency and reachability of the network when it is not attacked, respectively; and NS denotes the network stability (robustness).

Random and deliberate attacks have demonstrated significant applications in enhancing the resilience of networks. A comparative analysis of random and deliberate attacks can be employed in accident analysis to assess the performance of the network of accident-RIFs and validate the efficacy of the importance ranking of these RIFs. In random attacks, failure nodes were generated randomly following a uniform distribution, with the count of failure nodes incrementing gradually from 1 to encompassing all nodes. The CN robustness metric is derived by executing the random attack experiment 1000 times. For deliberate attacks, nodes are systematically targeted in descending order of importance based on the node ranking defined in Equation (16). The alterations in network robustness at each node failure were subsequently documented.

4. Results and discussions

4.1 Association rule mining results

To establish connections between the RIFs, the data underwent initial analysis using association rules. The determination of the ARM threshold significantly impacts the quality of the rules generated (Fu *et al.*, 2022; Zhou *et al.*, 2019b). This study employed a trial-and-error approach to find suitable thresholds. Initially, two sets of thresholds, with minimum support and minimum confidence set at 0.01 and 0.1 respectively, and 0.2 and 0.5 respectively, were selected to determine the association rules of ship collision accidents, based on relevant literature (Fu *et al.*, 2022; Huang *et al.*, 2023). When using the first set of thresholds, 1742 association rules were discovered. However, with the second set of thresholds, only 61 association rules were found. Following these two experiments, it was observed that these results did not provide appropriate association rules, making it difficult to explore the interactions between RIFs. To ensure the robustness and effectiveness of the rules, after multiple experiments, this study ultimately set the minimum support threshold and the minimum confidence threshold to 0.1 and 0.3 respectively. Additionally, to ensure that each node in the constructed CN represents only one RIF, the maximum limit length was set to 2 (Lan *et al.*, 2023).

Compared with the traditional Apriori algorithm, FP-growth algorithm is more efficient and more suitable for large sample analysis due to the use of FP-tree. Each of the two algorithms was run 10 times with 100 cycles per run, and the running time was shown in Table 2. It was then, found that the running efficiency of FP-Growth is 3.2 times that of Apriori. Therefore, FP-Growth algorithm was chosen to generate 276 association rules in this study, among which the top 10 association rules with confidence are shown in Table 3. It can be seen that four of these rules are related to management factors, which indicates that there is a strong correlation among various management factors and that they have an important impact on marine accidents. In addition, H2 (Insufficient education and training) and H3 (Lack of experience) exhibit the most pronounced elevation, indicating a substantial surge in the likelihood of crew inexperience when confronted with inadequate crew education and training. Insufficient educational preparation implies that crew members may not have acquired the requisite knowledge and skills essential for maritime tasks and emergency response. Consequently, this deficiency may result in crew members possessing inadequate familiarity with the maritime domain, ship operations, and emergency procedures. It also implies a potential lack of opportunities for crew members to apply theoretical knowledge in controlled settings, thereby impeding the acquisition of essential practical experience. The deficiency in education and training thus directly impacts the practical experience opportunities for crew members, giving rise to inexperience. This positive correlation underscores the potential contribution of enhanced education and training levels to the professionalism of crew members, fostering improved adaptation to intricate maritime tasks and challenging situations.

Table 2 Comparison of operation efficiency between FP-Growth and Apriori.

Algorithm	FP-Growth	Apriori
Time	$154ms \pm 4.38ms$	$493ms \pm 849\mu s$

Table 3 Top 10 rules ranked by confidence values.

NO.	Antecedent	Consequent	Support	Confidence	Lift
1	SG3	SE3	0.50	0.97	1.80
2	SE1	SG1	0.30	0.96	2.68
3	M7	M6	0.28	0.93	2.86
4	M5	M3	0.28	0.93	1.90
5	SE3	SG3	0.50	0.93	1.80
6	ST7	SG1	0.15	0.92	2.57
7	ST7	SE1	0.14	0.91	2.90
8	H2	H3	0.10	0.91	3.59
9	M2	M3	0.29	0.87	1.78

10	M6	M7	0.28	0.86	2.86

4.2 Results of weighted directed complex network

4.2.1 Topology properties of the weighted directed complex network

Utilizing the outcomes of ARM, this study established a directed CN comprising 31 nodes and 276 edges, as illustrated in Figure 6. The topological characteristics of the network were scrutinized to unveil the interactive nature of RIFs in marine accidents. The network's ACC value is 0.651, markedly surpassing the value of ACC (0.297) observed in a random network of equivalent size, for which the relevant parameter for a random network of equivalent size is computed as the average of the parameters obtained from 1000 randomly generated networks using the python code. This discrepancy underscores the tightly interconnected structure of the present network, implying that the RIFs within it exhibit a pronounced propensity to mutually influence each other. Additionally, the ASPL for this network is 1.469, notably lower than the 1.749 observed in a random network of equivalent size. This disparity implies a rapid establishment of connections between any two non-adjacent nodes within the network. Such findings suggest the possibility of indirect interactions among RIFs that may lack direct cross-influence, mediated through concise chain reactions. So that, the comprehensive network of RIFs associated with marine accidents exhibits greater cohesion, and the swift dissemination of information adds complexity to the challenge of accident reduction.

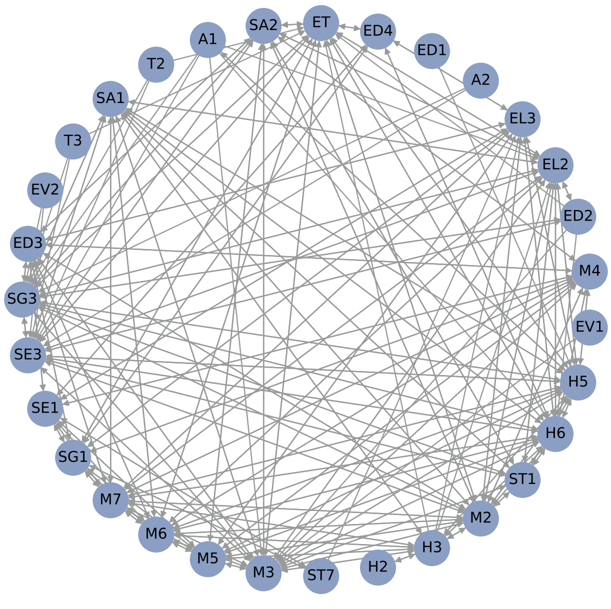

Figure 6 Interaction network of RIFs of marine accidents.

(1) Node degree and strength

The node degree signifies the various potential interactions among nodes and quantifies the direct connectivity of a node with other nodes in the network. Illustrated in Figure 7, nodes exhibit varying levels of activity when viewed through the lens of node degree. Notably, M3 (Defective safety management system) boasts the highest total degree value, closely trailed by H6 (Violation rules and regulations), M2 (Inadequate supervision), and H5 (Operational error). A higher total degree value signifies a greater number of interacting nodes. Additionally, it was observed that M2 (Inadequate supervision), H5 (Operational errors), H6 (Violation rules and regulations), and M3 (Defective safety management system) exhibit the highest out-degree values, equal to or exceeding 16, signifying their capacity to interact with at least 50% of the nodes in the network. This implies a wide-ranging influence of these RIFs within the marine accident system. Conversely, nodes with high in-degree values include M3 (Defective safety management system), H6 (Violation of rules and regulations), H5 (Operational error), SG3 (Gross tonnage ≥ 3000t), M2 (Inadequate supervision), SE3 (Engine power ≥ 3000KW), EL2 (Port), and ET (Heavy traffic), indicating that these RIFs are significantly influenced by interactions with a majority of other RIFs. Nodes exhibiting a degree value of 0, including EV1 (Poor visibility, 0.5-2 nm), EV2 (Very poor visibility, <0.5 nm), ED1 (h/d <1.2), A2 (Stranding/Ground), T2 (0400-0800), and T3 (0800-1200), indicate that these specific RIFs are not influenced by other RIFs while exerting an influence on other RIFs.

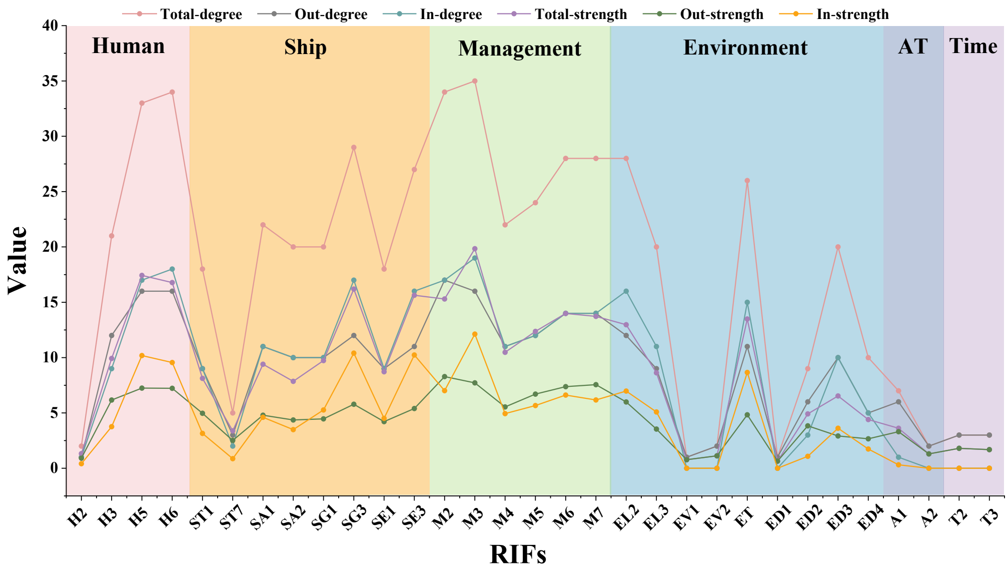

Figure 7 Degree and Strong degree distribution of CN.

While the node degree provides a measure of the direct interactions a node can have with other nodes, it does not account for the probability and intensity of these interactions. Examining node strength offers a different perspective, revealing instances where nodes may interact with a substantial number of other nodes, yet the probability and intensity of these interactions remain relatively low. For instance, although the node degree of EL2 (Port) surpasses that of ET (Heavy traffic), the strength of ET is higher, indicating a higher likelihood of interaction between ET and other nodes, despite the fewer nodes directly interacting with it. Similarly, M7 (Not drill by sticking to the schedule) exhibits a smaller out-degree compared to H6 (Violation rules and regulations), but it’s out-strength is notably larger, suggesting that M7 possesses a relatively higher probability of interactive influence on other nodes and exerts greater influence within the network.

Certainly, it is discernible that certain influences exhibit higher magnitudes than others, both in terms of degree and intensity. Notably, the parameters associated with ET (Heavy traffic) consistently manifest higher values compared to those linked with EV1 (Poor visibility, 0.5-2 nm) and EV2 (Very poor visibility, <0.5 nm). This implies that while visibility has a broad impact on marine accidents (Rawson *et al.*, 2021), the influence of traffic congestion within the accident system may exert a more pronounced effect.

(2) Node betweenness centrality coefficient

The betweenness centrality coefficient serves as an indicator of the pivotal role a node plays in facilitating information transfer within a complex system. A higher betweenness centrality coefficient suggests that the node acts as a central point through which more information flows, positioning it at the network's core. In Figure 8, the distribution of node betweenness centrality coefficient is depicted. Notably, H5 (Operational error) stands out as the node with the highest betweenness centrality coefficient across the entire network. This observation implies that operational errors by the crew serve as a crucial bridge for information transfer and influence within the overarching accident risk network. It suggests that various RIFs may interact and propagate along the path of crew operational error, emphasizing the effectiveness of reducing crew operational accidents in impeding or decelerating the transfer of information between RIFs, consequently reducing or slowing down the occurrence of accidents.

On the contrary, H2 (Insufficient education and training), ST7 (Fishing vessel), EV1 (Poor visibility, 0.5-2 nm), EV2 (Very poor visibility, <0.5 nm), ED1 (h/d <1.2), ED2 (1.2<=h/d<1.5), A1 (Collision), A2 (Stranding/Ground), T2 (0400-0800), and T3 (0800-1200) all exhibit a betweenness centrality coefficient of 0. A betweenness centrality coefficient of zero signifies that these RIFs do not fulfill a mediating role in the network. In other words, they are not key RIFs connecting other RIFs and may not assume a significant bridging function in the accident risk network. This could stem from the fact that these RIFs lack an evident propagation path within the network or may possess weak correlations among themselves. Alternatively, a betweenness centrality coefficient of 0 might result from these RIFs playing a role only in specific contexts, lacking critical significance in other scenarios. For instance, time horizons (T2 and T3) may exclusively impact accident risk during specific time periods, and 'Fishing vessel' (ST7) might possess distinctive properties that exempt it from mediating in particular risk propagation pathways. Consequently, in mitigating accident risk, greater attention may be warranted for RIFs exhibiting higher betweenness centrality coefficient in the network, as they are more likely to assume a key connecting role in risk propagation.

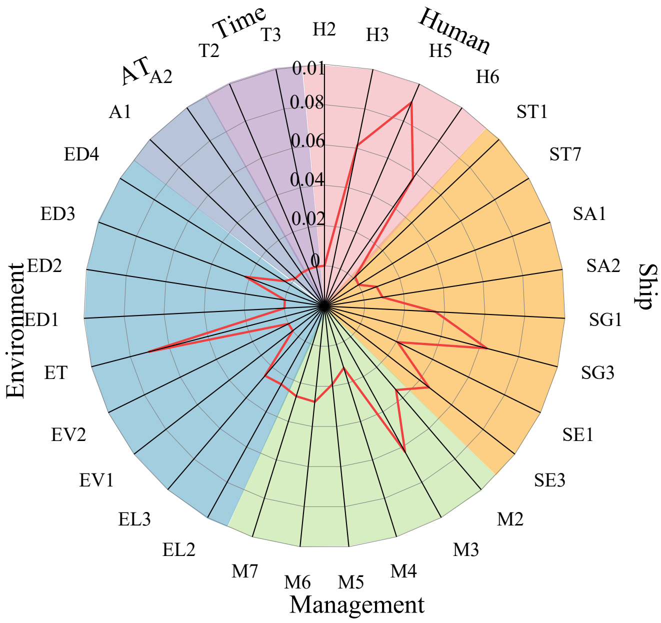

Figure 8 Betweenness centrality coefficient distribution of CN.

4.2.2 Node importance

The determination of the significance of nodes in the influencing interaction network employed the WLR algorithm. Establishing a suitable parameter, denoted as α, is of paramount importance, as it exerts a certain influence on the convergence (iteration number) of the algorithm, albeit having minimal impact on the ranking order of node importance (Li *et al.*, 2014). Thus, in this study, an optimal parameter α was identified through a series of comparative experimental validations. The efficiency of iteration convergence under different scenarios was expressed as the ratio between the convergence speed of each iteration when $\alpha = 1$ and that of each iteration with varying α values, as illustrated in Figure 9.

The finalized ranking of node importance is detailed in Table 4. Notably, M2 (Inadequate supervision) emerges as the most crucial, followed by M3 (Defective safety management system) and H6 (Violation rules and regulations). Conversely, EV1 (Poor visibility, 0.5-2 nm), A2 (Stranding/Ground), ED1 (h/d < 1.2), T2 (0400-0800), T3 (0800-1200), and EV2 (Very poor visibility, <0.5 nm) exhibit the least significance.

The results underscore the profound impact of insufficient regulation, establishing it as a pivotal RIFs of marine accidents. Inadequate regulatory and supervisory mechanisms may lead to inappropriate and unsafe behaviors among ships and maritime traffic participants. Enhancing regulatory frameworks and oversight is posited to play a pivotal role in reducing accident risks. Similarly, defective safety management systems and violations of regulations emerge as pivotal RIFs influencing accident risks. This could entail deficiencies in the safety management systems of shipping companies or affiliated organizations, coupled with breaches of regulations by crew members or other maritime traffic participants. Strengthening safety management systems and enforcing regulations holds promise for mitigating accident risks.

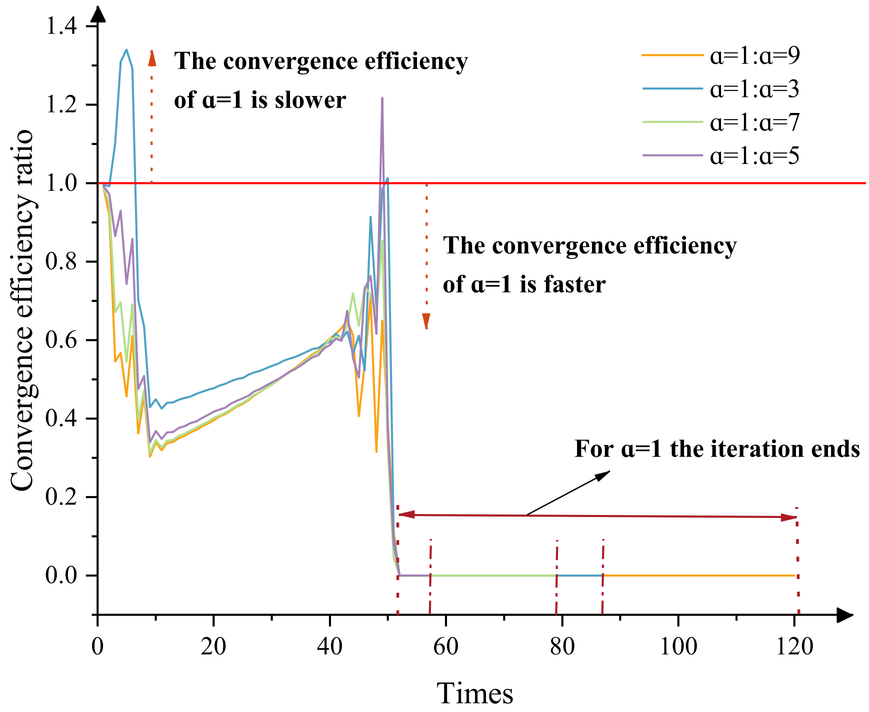

Figure 9 Convergence efficiency ratio of $\alpha = 1$ to other α .

Table 4 Nodes importance ranking.

Rank	RIF	WLR	Rank	RIF	WLR	Rank	RIF	WLR
1	M2	2.06	12	M4	1.35	23	ST7	0.29
2	M3	2.02	13	SA1	1.34	24	A1	0.19
3	H6	2.01	14	SA2	1.27	25	H2	0.16
4	H5	1.87	15	EL3	1.25	26	EV1	0.07
5	M6	1.68	16	ED3	1.24	27	A2	0.07
6	M7	1.67	17	SG1	1.18	28	ED1	0.07
7	EL2	1.67	18	ST1	1.15	29	T2	0.07
8	ET	1.53	19	H3	1.07	30	T3	0.07
9	SG3	1.51	20	SE1	1.06	31	EV2	0.07
10	M5	1.45	21	ED4	0.67			

RIFs such as EV1 (Poor visibility, 0.5-2 nm), A2 (Stranding/Ground), ED1 (h/d < 1.2), T2 (0400-0800), T3 (0800-1200), and EV2 (Very poor visibility, <0.5 nm) are regarded as less influential. While their contribution to the overall risk in the current network structure appears relatively limited, it is crucial to note that their lower position in the current network does not diminish their inherent importance. This ranking offers valuable insights for shaping comprehensive strategies and interventions. Prioritizing high-importance RIFs, particularly those related to regulatory issues, safety management systems, and adherence to rules and regulations, stands out as an effective approach to enhance maritime safety.

4.3 Robustness analysis of the weighted directed complex network

The robustness of a network refers to its capacity to maintain stability in the face of node or edge failures. Disruption to the marine accidents' association network of RIFs, and the subsequent reduction in network stability, can mitigate the evolution of accidents to a certain extent. This study assesses network robustness by employing stability metrics (as Equation (20)), random attacks, deliberate attacks based on PageRank node importance ranking (Meng *et al.*, 2022), and deliberate attacks based on LeaderRank algorithm node importance ranking (Zhou *et al.*, 2019a). These are compared with deliberate attacks based on the WLR algorithm node importance ranking to underscore the validity of the WLR algorithm's ranking outcomes. The simulation results depicting network robustness are illustrated in Figure 10.

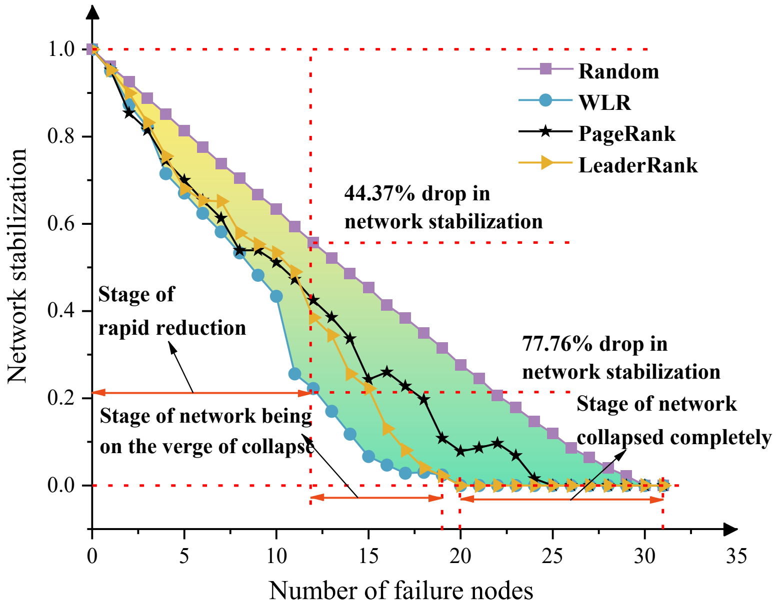

Figure 10 Network robustness simulation results.

In Figure 10, it is evident that among the three deliberate attacks, the deliberate attack utilizing the WLR algorithm exhibits a faster network collapse, suggesting a higher degree of correctness and effectiveness in the sorting results of the WLR algorithm. Furthermore, a comparison between deliberate attacks and random attacks based on the importance ranking of WLR nodes reveals that the impact on network stability varies with the increasing number of deleted nodes. Specifically, the network stability experiences a slower decrease in the case of random attacks compared to deliberate attacks. For instance, a random attack on 12 nodes results in a 44.37% decrease in network stability, while a deliberate attack on the same number of nodes leads to a more substantial decrease of 77.76%. The changing area between the curves representing random and deliberate attacks indicates that when the total number of attack nodes is less than 16, the disparity in the influence on network stability between deliberate and random attacks becomes more pronounced with an increasing number of failed nodes. This observation underscores that deliberate attacks on critical nodes can induce rapid changes in the network structure. Contrastingly, in random attacks, the probability of targeting key nodes is relatively low. It is only when a large number of nodes are subjected to attacks that substantial alterations in the network structure occur, potentially leading to network collapse. Figure 10 illustrates that the network stability is a mere 17.00% when the 13th node is deliberately attacked, indicating a stage of rapid decline and proximity to collapse. As deliberate attacks progress and reach the 20th node, the network stability plunges to 0, signifying a complete collapse. This emphasizes the efficacy of purposefully targeting fewer critical nodes in accident prevention and control to achieve the objective of preventing or managing accidents.

5. Implication

Based on the results of analyzing the topological characteristics of the interaction network of RIFs in marine accidents and the importance of the nodes, this section proposes counter-measures for the key RIFs from four aspects: human factors, ship factors, management factors, and environmental factors. The aim is to effectively prevent and control these RIFs, thereby reducing accident rates and ensuring the safety and reliability of maritime transportation.

(1) Among the human factors, H6 (Violation of rules and regulations), H5 (Operational error) are the more important RIFs. Therefore, regularly conduct skill testing and training for crew members, and provide skill upgrading training for unqualified personnel, to ensure their familiarity with and mastery of general marine equipment, thereby enabling effective response to emergency situations. Additionally, it is suggested to establish an incentive mechanism to encourage crew members to actively participate in training activities, fostering continuous improvement of their personal attributes. Furthermore, crew members should thoroughly study ship and route information in advance, gaining a comprehensive understanding of the vessel's navigation characteristics and the environmental conditions along the route, in order to avoid erroneous decisions resulting from insufficient knowledge. Therefore, it is crucial to select crew members with wide navigational experience to ensure correct maneuvering and operation of the vessel, preventing errors due to lack of vessel familiarity.

(2) Within the realm of ship factors, SG3 (Gross tonnage ≥ 3000t) and SE3 (Engine power ≥ 3000KW) emerge as pivotal RIFs, underscoring the significance of larger ship. Consequently, regular inspections and maintenance protocols should be implemented to uphold the operational integrity of these sizable ships. Furthermore, there is a pressing need to enhance the training programs for crew members, focusing specifically on the operation and emergency management procedures relevant to large ships, thus bolstering their proficiency in handling crises effectively. Lastly, comprehensive risk assessments, particularly targeting the mechanical components of large ships, are imperative to proactively identify potential failures and rectify underlying issues.

(3) Among the management factors, M3 (Defective safety management system) and M2 (Inadequate supervision) emerge as the more critical RIFs. Hence, it is imperative to enhance and proficiently execute a ship safety management system. This system ought to encompass standardized procedures for inspection, record-keeping, and reporting, alongside a structured framework for incident response. Concurrently, it is advisable to establish a specialized supervisory team dedicated to ensuring the effective implementation of

these measures. Significantly, emphasis should be placed on bolstering equipment maintenance and management practices to guarantee the safe and reliable operation of onboard equipment. Furthermore, crew members' operational and emergency response capabilities should undergo regular and thorough assessment, while ship manning levels should be subject to ongoing monitoring.

(4) Among the environmental factors, EL2 (Port) and ET (Heavy traffic) emerge as the more significant RIFs. Therefore, enhancing the frequency of surveillance and patrols within port areas could effectively ensure ship compliance with maritime regulations. Additionally, stringent navigation management measures, such as traffic separation schemes and speed restrictions, ought to be enforced in regions characterized by high traffic density. Furthermore, regular risk assessments of both port areas and high traffic density zones are essential, alongside the formulation of contingency plans to promptly adapt preventive measures as needed.

6. Conclusions

This study introduces a robust analytical framework for investigating the interactions among RIFs in marine accidents. The strength of this research lies in its integration of ARM and CN, which not only identifies the frequent itemsets of RIFs leading to marine accidents but also elucidates the interconnections between these RIFs and the risk evolution process. ARM offers valuable insights by extracting common association relationships among RIFs from historical accident reports in marine settings. These relationships are subsequently mapped onto a CN to visualize the interaction of marine accident RIFs, providing a comprehensive articulation and interpretation of the complex interactions among marine risks and the pivotal role of RIFs in the evolutionary process. This approach effectively delineates the position of RIFs within the marine accident system, thus highlighting their significance as key contributors. These insights serve as a foundation for the prevention of marine accidents and the enhancement of maritime safety.

The results reveal that the RIFs association network in marine accidents is more densely interconnected than a random network, with faster information transfer speeds. This underscores the challenges associated with controlling marine accidents. Furthermore, insufficient education and training significantly amplify crew inexperience, underscoring the necessity of improving crew education and training. Addressing these aspects enhances crew adaptation to complex maritime environments and bolsters their ability to handle emergencies, ultimately reducing the occurrence of marine accidents. Crew operational errors emerge as pivotal conduits for information transfer and impact within the marine accident system, emphasizing the potential positive impact of enhancing crew operational standards and reducing errors. Strategies involving improvements in crew training and operating procedures to minimize human error and operational mistakes are recognized as crucial for enhancing the effectiveness of marine traffic accident prevention and control. Notably, inadequate regulation, crew non-compliance, and deficiencies in safety management systems emerge as the most significant RIFs, underscoring the urgency of regulatory improvement, strengthened enforcement, and enhanced safety management systems.

While this study contributes significantly by unveiling the risk interactions among marine accidents and identifying key RIFs, it also entails certain limitations warranting future consideration. Firstly, the construction of the dataset involved subjective inputs such as categorization. Future research could enhance dataset construction by employing deep learning-based text mining techniques. Secondly, although ARM techniques effectively analyze association relationships between RIFs, such association do not imply causation. To delve deeper into the causal role relationships of each RIF in marine accidents, future attempts could involve advanced causal discovery algorithms. Moreover, this study simplifies the model by assigning each node to represent only one RIF, without considering cases where a node represents a collection of multiple RIFs. Lastly, this study overlooks the effect of accident severity during the model construction process. Subsequent studies could further explore these findings to analyze the marine accident system in greater detail, incorporating considerations of accident severity.

Data Availability Statement

All data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration of Competing Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and publication of this article.

Acknowledgments

The authors gratefully acknowledge support from the National Natural Science Foundation of China (Grant No. 52101399), Bolian Research Funds of Dalian Maritime University (Grant No. 3132023617). This work was also supported by the EU H2020 ERC Consolidator Grant program (TRUST Grant No. 864724).

Appendix

Table A Description of the 68 RIFs

RIFs		Variables	Description
Human		H1	Poor state of physical & psychological
		H2	Insufficient education and training
		H3	Lack of experience
		H4	Poor communication
		H5	Operational error
		H6	Violation rules and regulations
Ship	Ship type	ST1	Bulk carrier
		ST2	Container ship
		ST3	Oil tanker
		ST4	Passenger ship (including cruise and ro-ro passenger ship)
		ST5	Chemical tanker
		ST6	General cargo ship
		ST7	Fishing vessel
		ST8	Yacht and sailing vessel
		ST9	Tug and port traffic boat
		ST10	Others
	Ship age	SA1	0-10years
		SA2	10-20years

		SA3	20-30years
		SA4	>=30 years
	Gross tonnage	SG1	<500t
		SG2	500-3000t
		SG3	>=3000t
	Engine power	SE1	<750KW
		SE2	750-3000KW
		SE3	>=3000KW
	Voyage data	SV1	Incomplete or invalid ship certificates
		SV2	Insufficient manning
		SV3	Incomplete or invalid seafarer certificates
		SV4	Poor seaworthiness
		SV5	Uncertain PSC/FSC inspection
Management		M1	Inadequate regulation
		M2	Inadequate supervision
		M3	Defective safety management system
		M4	Unresponsive rectification of problems
		M5	Poor company safety culture
		M6	Inadequate ship training
		M7	Not drill by sticking to the schedule
Environment	Location	EL1	Inland waters
		EL2	Port
		EL3	Coastal waters
		EL4	Open sea
	Visibility	EV1	Poor visibility, 0.5-2nm
		EV2	Very Poor visibility, <0.5nm
	Wind force	EW1	Strong wind, 6-7
		EW2	Very strong wind, 8-9
		EW3	Strom, 10-12
	Sea State	ES1	Poor sea state, 6-7
		ES2	Very poor sea state, 8-9
	Current	EC	Current speed >=4kn
	Traffic density	ET	Heavy traffic
	Depth-draft ratio (h/d)	ED1	h/d < 1.2
		ED2	1.2 <= h/d < 1.5
		ED3	1.5 <= h/d < 3
		ED4	h/d >= 3
Accident	Accident type (AT)	A1	Collision
		A2	Stranding/Ground
		A3	Fire/Explosion
		A4	Contact
		A5	Capsize/Sinking
		A6	Hull/Machinery damage
		A7	Other
	Time	T1	0000-0400
		T2	0400-0800
		T3	0800-1200
		T4	1200-1600
		T5	1600-2000
		T6	2000-2400

Note: The unit for wind force is "Beaufort scale"; the unit for sea state is "Douglas Sea Scale."

Reference

AGCS, 2023. Safety and Shipping Review 2023. Allianz Global Corporate & Specialty

Çakır, E., Fışkın, R., Sevgili, C., 2021. Investigation of tugboat accidents severity: An application of association rule mining algorithms. *Reliability Engineering & System Safety* 209, 107470.https://doi.org/10.1016/j.ress.2021.107470

Feng, Y., Wang, H., Xia, G., Cao, W., Li, T., Wang, X., Liu, Z., 2025. A machine learning-based data-driven method for risk analysis of marine accidents. *Journal of Marine Engineering & Technology*, 24 (2), 147-158. https://doi.org/10.1080/20464177.2024.2368914

Cakir, E., Sevgili, C., Fiskin, R., 2021. An analysis of severity of oil spill caused by vessel accidents. *Transportation Research Part D: Transport and Environment* 90, 102662.https://doi.org/10.1016/j.trd.2020.102662

Cao, Y., Wang, X., Wang, Y., Fan, S., Wang, H., Yang, Z., Liu, Z., Wang, J., Shi, R., 2023a. Analysis of factors affecting the severity of marine accidents using a data-driven Bayesian network. *Ocean Engineering* 269, 113563.https://doi.org/10.1016/j.oceaneng.2022.113563

Cao, Y., Wang, X., Yang, Z., Wang, J., Wang, H., Liu, Z., 2023b. Research in marine accidents: A bibliometric analysis, systematic review and future directions. *Ocean Engineering* 284, 115048.https://doi.org/10.1016/j.oceaneng.2023.115048

Christensen, M., Georgati, M., Arsanjani, J.J., 2022. A risk-based approach for determining the future potential of commercial shipping in the Arctic. *Journal of Marine Engineering & Technology* 21(2), 82-99.https://doi.org/10.1080/20464177.2019.1672419

EMSA, 2022. Annual Overview of Marine Casualties and Incidents 2022. European Maritime Safety Agency, Lisbon, Portugal, p. 64

Fan, S., Yang, Z., 2024. Accident data-driven human fatigue analysis in maritime transport using machine learning. *Reliability Engineering & System Safety* 241, 109675.https://doi.org/10.1016/j.ress.2023.109675

Fang, S., Liu, Z., Wang, X., Cao, Y., Yang, Z., 2024. Dynamic analysis of emergency evacuation in a rolling passenger ship using a two-layer social force model. *Expert Systems with Applications* 247, 123310.https://doi.org/10.1016/j.eswa.2024.123310

Feng, Y., Wang, Z., Dong, Q., 2021. How to Analyze the Urban Rail Transit Network in the Combination View of Complex Network and Clustering: Take Shenzhen as an Example, *2021 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pp. 624-631.http://10.1109/AINIT54228.2021.00126

Fu, L., Wang, X., Zhao, H., Li, M., 2022. Interactions among safety risks in metro deep foundation pit projects: An association rule mining-based modeling framework. *Reliability Engineering & System Safety* 221, 108381.https://doi.org/10.1016/j.ress.2022.108381

Huang, C.H., Hu, S.P., 2019. Factors correlation mining on maritime accidents database using association rule learning algorithm. *CLUSTER COMPUTING-THE JOURNAL OF NETWORKS SOFTWARE TOOLS AND APPLICATIONS* 22(2), S4551-S4559.http://10.1007/s10586-018-2089-z

Huang, D., Liang, T., Hu, S., Loughney, S., Wang, J., 2023. Characteristics analysis of intercontinental sea accidents using weighted association rule mining: Evidence from the Mediterranean Sea and Black Sea. *Ocean Engineering* 287, 115839.https://doi.org/10.1016/j.oceaneng.2023.115839

Hwang, T., Youn, I.-H., 2022. Latent-Cause Extraction Model in Maritime Collision Accidents Using Text Analytics on Korean Maritime Accident Verdicts. *Applied Sciences* 12(2), 914.http://10.3390/app12020914

IMO, 2023. Maritime Safety, access on 30th April, 2024. https://www.imo.org/en/OurWork/Safety/Pages/default.aspx

Kimera, D., Nangolo, F.N., 2022. Reliability maintenance aspects of deck machinery for ageing/aged fishing vessels. *Journal of Marine Engineering & Technology* 21(2), 100-110.https://doi.org/10.1080/20464177.2019.1663595

Lan, H., Ma, X., Qiao, W., Deng, W., 2023. Determining the critical risk factors for predicting the severity of ship collision accidents using a data-driven approach. *Reliability Engineering & System Safety* 230, 108934.https://doi.org/10.1016/j.ress.2022.108934

Lee, J.-S., Yu, Y.-U., 2023. Calculation of categorical route width according to maritime traffic flow data in the Republic of Korea. *Journal of Marine Engineering & Technology* 22(5), 222-232.https://doi.org/10.1080/20464177.2023.2223396

Li, H., Jia, P., Wang, X., Yang, Z., Wang, J., Kuang, H., 2023a. Ship carbon dioxide emission estimation in coastal domestic emission control areas using high spatial-temporal resolution data: A China case. *Ocean & Coastal Management* 232, 106419.https://doi.org/10.1016/j.ocecoaman.2022.106419

Li, H., Ren, X., Yang, Z., 2023b. Data-driven Bayesian network for risk analysis of global maritime accidents. *Reliability Engineering & System Safety* 230, 108938.https://doi.org/10.1016/j.ress.2022.108938

Li, Q., Zhou, T., Lu, L.Y., Chen, D.B., 2014. Identifying influential spreaders by weighted LeaderRank. *PHYSICA A-STATISTICAL MECHANICS AND ITS APPLICATIONS* 404, 47-55.http://10.1016/j.physa.2014.02.041

Meng, X., Han, S., Wu, L., Si, S., Cai, Z., 2022. Analysis of epidemic vaccination strategies by node importance and evolutionary game on complex networks. *Reliability Engineering & System Safety* 219, 108256.https://doi.org/10.1016/j.ress.2021.108256

Narvekar, M., Syed, S.F., 2015. An Optimized Algorithm for Association Rule Mining Using FP Tree. *Procedia Computer Science* 45, 101-110.https://doi.org/10.1016/j.procs.2015.03.097

Nguyen, T., Zhang, J., Zhou, L., Meng, M., He, Y., 2019. A data-driven optimization of large-scale dry port location using the hybrid approach of data mining and complex network theory. *Transportation Research Part E: Logistics and Transportation Review* 134.http://10.1016/j.tre.2019.11.010

Rawson, A., Brito, M., Sabeur, Z., Tran-Thanh, L., 2021. A machine learning approach for monitoring ship safety in extreme weather events. *Safety Science* 141, 105336.https://doi.org/10.1016/j.ssci.2021.105336

Shafiee, M., Animah, I., 2022. An integrated FMEA and MCDA based risk management approach to support life extension of subsea facilities in high-pressure–high-temperature (HPHT) conditions. *Journal of Marine Engineering & Technology* 21(4), 189-204.https://doi.org/10.1080/20464177.2020.1827486

Shi, J., Liu, Z., Feng, Y., Wang, X., Zhu, H., Yang, Z., Wang, J., Wang, H., 2024. Evolutionary model and risk analysis of ship collision accidents based on complex networks and DEMATEL. *Ocean Engineering* 305, 117965.https://doi.org/10.1016/j.oceaneng.2024.117965

Solanki, S.K., Patel, J.T., 2015. A Survey on Association Rule Mining, *2015 Fifth International Conference on Advanced Computing & Communication Technologies*, pp. 212-216.http://10.1109/ACCT.2015.69

Feng, Y., Wang, H., Xia, G., Cao, W., Li, T., Wang, X., Liu, Z., 2025. A machine learning-based data-driven method for risk analysis of marine accidents. *Journal of Marine Engineering & Technology*, 24 (2), 147-158. https://doi.org/10.1080/20464177.2024.2368914

Tian, Y.-F., Chen, L.-J., Huang, L.-W., Mou, J.-M., 2020. Featured risk evaluation of nautical navigational environment using a risk cloud model. *Journal of Marine Engineering & Technology* 19(3), 115-129.http://10.1080/20464177.2018.1534929

Ugurlu, H., Cicek, I., 2022. Analysis and assessment of ship collision accidents using Fault Tree and Multiple Correspondence Analysis. *Ocean Engineering* 245, 110514.https://doi.org/10.1016/j.oceaneng.2021.110514

Valcalda, A., de Koningh, D., Kana, A., 2023. A method to assess the impact of safe return to port regulatory framework on passenger ships concept design. *Journal of Marine Engineering & Technology* 22(3), 111-122.https://doi.org/10.1080/20464177.2022.2031557

Wang, H., Liu, Z., Wang, X., Graham, T., Wang, J., 2021. An analysis of factors affecting the severity of marine accidents. *Reliability Engineering & System Safety* 210, 107513.https://doi.org/10.1016/j.ress.2021.107513

Wang, H., Liu, Z., Wang, X., Huang, D., Cao, L., Wang, J., 2022a. Analysis of the injury-severity outcomes of maritime accidents using a zero-inflated ordered probit model. *Ocean Engineering* 258, 111796.https://doi.org/10.1016/j.oceaneng.2022.111796

Wang, J., Wang, X., Feng, Y., Cao, Y., Guo, Z., Liu, Z., 2023a. Assessing the Connectivity Reliability of a Maritime Transport Network: A Case of Imported Crude Oil in China. *Journal of Marine Science and Engineering* 11(8), 1597.https://doi.org/10.3390/jmse11081597

Wang, J., Zhou, Y., Zhuang, L., Shi, L., Zhang, S., 2022b. Study on the critical factors and hot spots of crude oil tanker accidents. *Ocean & Coastal Management* 217, 106010.https://doi.org/10.1016/j.ocecoaman.2021.106010

Wang, J., Zhou, Y., Zhuang, L., Shi, L., Zhang, S., 2023b. A model of maritime accidents prediction based on multi-factor time series analysis. *Journal of Marine Engineering & Technology* 22(3), 153-165.https://doi.org/10.1080/20464177.2023.2167269

Wang, X., Liu, Z., Loughney, S., Yang, Z., Wang, Y., Wang, J., 2022c. Numerical analysis and staircase layout optimisation for a Ro-Ro passenger ship during emergency evacuation. *Reliability Engineering & System Safety* 217, 108056.https://doi.org/10.1016/j.ress.2021.108056

Weng, J., Liao, S., Wu, B., Yang, D., 2020. Exploring effects of ship traffic characteristics and environmental conditions on ship collision frequency. *Maritime Policy & Management* 47(4), 523-543.http://10.1080/03088839.2020.1721584

Xia, G., Wang, X., Feng, Y., Cao, Y., Dai, Z., Wang, H., Liu, Z., 2023. Navigational Risk of Inland Water Transportation: A Case Study in the Songhua River, China. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* 9(4), 04023042.http://10.1061/AJRUA6.RUENG-1158

Yu, Y., Chen, L., Shu, Y., Zhu, W., 2021. Evaluation model and management strategy for reducing pollution caused by ship collision in coastal waters. *Ocean & Coastal Management* 203, 105446.https://doi.org/10.1016/j.ocecoaman.2020.105446

Zhang, M., Conti, F., Le Sourne, H., Vassalos, D., Kujala, P., Lindroth, D., Hirdaris, S., 2021. A method for the direct assessment of ship collision damage and flooding risk in real conditions. *Ocean Engineering* 237, 109605.https://doi.org/10.1016/j.oceaneng.2021.109605

Zhang, M., Zhang, D., Fu, S., Kujala, P., Hirdaris, S., 2022. A predictive analytics method for maritime traffic flow complexity estimation in inland waterways. *Reliability Engineering & System Safety* 220, 108317.https://doi.org/10.1016/j.ress.2021.108317

Zhang, Z., He, Q., Gao, J., Ni, M., 2018. A deep learning approach for detecting traffic accidents from social media data. *Transportation Research Part C: Emerging Technologies* 86, 580-596.http://10.1016/j.trc.2017.11.027

Zhao, J., Yan, Z., Zhou, Z., Chen, X., Wu, B., Wang, S., 2023. A ship trajectory prediction method based on GAT and LSTM. *Ocean Engineering* 289, 116159.https://doi.org/10.1016/j.oceaneng.2023.116159

Zhou, B., Lei, Y., Li, C., Fang, B., Wu, Q., Li, L., Li, Z., 2019a. Electrical LeaderRank method for node importance evaluation of power grids considering uncertainties of renewable energy. *International Journal of Electrical Power & Energy Systems* 106, 45-55.https://doi.org/10.1016/j.ijepes.2018.09.034

Zhou, Y., Li, C., Ding, L., Sekula, P., Love, P.E.D., Zhou, C., 2019b. Combining association rules mining with complex networks to monitor coupled risks. *Reliability Engineering & System Safety* 186, 194-208.https://doi.org/10.1016/j.ress.2019.02.013