A Novel Multi-task Causal Representation Learning Approach for Interpretable Maritime Collision Severity Prediction

Miaomiao Wang[a], Yanfu Wang[a,b,c *], Zhicheng Ma[a], Jin Wang[c]

[a] Department of Safety Science and Engineering, College of Mechanical and Electronic Engineering, China University of Petroleum, Qingdao 266580, P.R. China;
[b] State Key Laboratory of Chemical Safety, China University of Petroleum (East China), Qingdao 266580, P.R. China;
[c] Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool UK
* Corresponding author, E-mail: wangyanfu@upc.edu.cn

Abstract: Accurate prediction and robust interpretation of maritime collision severity are crucial. Prevailing correlation-based methods are non-robust and lack interpretability, struggling with confounding factors, data heterogeneity, and class imbalance. A novel multi-task causal learning (MCLF) framework is proposed to address these limitations. Its core is a structured disentanglement mechanism that separates an influence into direct effects and indirect effects mediated by latent unsafe factors, reinforced by adversarial training and orthogonality constraints to mitigate confounding bias. To address class imbalance, an interactive data synthesis module using the tabular denoising diffusion probabilistic model (TabDDPM) is used, which generates high-quality samples for hard-to-classify cases to enhance model robustness. A dynamic multi-task fusion strategy then adaptively integrates the primary severity prediction with auxiliary tasks—pollution, property loss, and death. This holistic approach achieves superior predictive accuracy and enhances interpretability through a granular, cause-effect analysis, advancing reliable and transparent decision-support in maritime safety.

Keywords: Collision severity, Unsafe factors, Causal disentanglement, Multi-task fusion.

1. Introduction

As global maritime traffic density increases, the frequency of ship collision accidents rises accordingly. Data statistics from the Global Integrated Shipping Information System (GISIS) database, which contains maritime accident information reported to the International Maritime Organization (IMO), show that ship collisions account for about 20% of all maritime accidents. Ship collisions may not only cause huge economic losses, such as hull damage, cargo damage, and route interruption, and may also cause a series of chain reactions such as casualties, marine pollution (such as

oil leakage, spread of dangerous goods), and ecological damage. Especially in collisions involving high-tonnage, high-load merchant ships or tankers, the consequences of the accident are often more serious, with a wide range of impacts and extremely high recovery costs [1, 2]. In this context, understanding the underlying principles of how causal characteristics affect the severity of ship collisions, as well as understanding the generation mechanism and consequence patterns of ship collision accidents, is of great significance for predicting the severity of potential accidents, optimizing collision avoidance strategies and accident response mechanisms.

Causal analysis serves as a fundamental pillar of maritime safety, traditionally employed for the retrospective attribution of root and contributory causes to inform preventive strategies [3]. However, this paper shifts its focus from merely investigating the causes of accidents [4-7] to an in-depth examination of the complex causal pathways that determine the severity of consequences in ship collisions. Prevailing data-driven analytical frameworks often treat the evolution of an accident as a black box, thereby obscuring the critical intermediate mechanisms at play. Most existing studies [8, 9] regard unsafe factors such as human error, inadequate lookout, or potential equipment failure as observable input variables. Yet, in practice, these factors are typically unavailable before an accident, since they can only be identified through post-hoc investigations. This raises a critical question: Do external conditions like weather directly cause severe outcomes, or do they act indirectly by inducing unsafe mediating states (e.g., human error)? Overlooking these indirect pathways leads to misinterpreting the true impact of initial conditions. Therefore, disentangling direct from indirect causal effects is crucial. It enhances both predictive accuracy and interpretability by explicitly modeling how external factors trigger mediating risks to determine accident severity. While a recent study [10] has demonstrated the feasibility of inferring accident causes from observational features using advanced models, this task remains fundamentally retrospective in nature. Integrating such causal insights into pre-emptive severity prediction models remains a critical and unresolved challenge.

Traditional models are limited by a homogeneous-effect assumption, yielding a single, average causal narrative that masks the significant heterogeneity across diverse accident scenarios. In reality, the dominant causal pathways and their importance vary dramatically from one event to another. [11]. Furthermore, accident consequences are multi-dimensional (e.g., casualties, property loss and pollution), and models that predict these outcomes in isolation fail to capture their shared causal drivers, leading to

fragmented insights. Therefore, a new approach is needed to quantify instance-specific causal pathways while simultaneously modeling multiple, interrelated outcomes.

To address these limitations, we propose an interpretable multi-task framework grounded in causal representation learning, applied to a dataset of 686 ship collision accidents. The framework's key innovation is its use of causal disentanglement to perform path-specific analysis. This allows us to decompose the total effect of any risk factor into its direct impact and its indirect effects mediated through other unsafe conditions. By moving beyond average effects, our model provides a granular, quantitative understanding of the causal mechanisms in specific accidents, laying a robust foundation for more targeted and effective safety interventions.

The main contributions of this study are as follows:

- Interactive training mechanism between generation and classification models: we propose a feedback-driven joint optimization strategy that integrates a data generation model (TabDDPM) with a classification model. By adaptively generating synthetic samples near the decision boundary and feeding them back into the classifier, the framework directs learning toward difficult cases.
- Structural causal modeling for multi-task prediction: The integration of causal decoupling representation learning with multi-task optimisation mechanisms not only enhances predictive accuracy and generalisation capabilities but also enables systematic modelling of the intricate causal pathways underlying multi-dimensional accident consequences.
- Dynamic causal disentanglement mechanism for modeling causal heterogeneity: By using cross-attention methods with learnable causal queries to dynamically infer the most prominent potential unsafe states within each instance before an incident, the analytical paradigm fundamentally shifts from estimating homogeneous average effects to modelling heterogeneous individual effects.

The remainder of this study is organized as follows: Section 2 presents a review of related literature, followed by a detailed explanation of the proposed methodology in Section 3. The experimental results and corresponding discussions are provided in Section 4, and the conclusions are summarized in Section 5.

2. Literature reviews

2.1 Related research on maritime accident severity

Accidents rarely stem from a single or isolated human error or equipment failure, but rather arise from the complex interplay of multiple factors [12, 13], including personnel performance, equipment condition, environmental circumstances, and management procedures. Consequently, researchers are increasingly adopting a systems-analysis perspective to construct accident causal models. For example, Fu et al. [14] employed the AcciMap framework to identify potential risk factors and their interrelationships in ship grounding accidents under Arctic navigation conditions. Similarly, Ma et al. [15] integrated HFACS (Human Factors Analysis and Classification System), DEMATEL (Decision-Making Trial and Evaluation Laboratory), and FCM (Fuzzy Cognitive Map) to model the human factors contributing to maritime collisions. While these approaches have yielded valuable insights into risk identification, they primarily remain at the level of qualitative analysis and lack quantitative predictive and reasoning capabilities.

In parallel, researchers have attempted to extract co-occurrence rules from accident data, identify key factors as model inputs, and construct more sophisticated predictive models using methods such as association rule mining (ARM), complex networks (CN), and knowledge graphs. For instance, He et al. [9] integrated ARM, CN, and random forests (RF) to examine correlations among risk factors and to identify critical predictors of ship collision severity, highlighting poor team communication as the most influential risk factor. Chen et al. [16] developed a joint triple-extraction algorithm based on deep learning and CART (classification and regression tree) to construct accident knowledge graphs, and further applied complex network analysis to define safety-related topological indicators for quantitative risk assessment and the identification of key factors. Nevertheless, these methods are essentially grounded in statistical correlations rather than causal inference.

In addition, a range of machine learning techniques such as Bayesian networks [5, 13], support vector machines (SVM) [13] and random forests (RF) [9] have been widely applied to predict the severity of ship collisions. Lan et al. [9] combined ARM with CN to identify accident-related risk factors and then developed an RF model for severity prediction. Qiao et al. [2] proposed a data-driven hybrid algorithm (BiLSTM-CNN-RF) to capture latent patterns in the relationship between risk-inducing factors (RIF) and accident consequences in the maritime transport domain. Although these models

incorporate multiple variables for prediction, they are constrained by their single-task learning design. These methods typically employ multiple factors such as vessel characteristics, waterway features and environmental conditions to predict isolated outcomes (accident severity ratings), failing to comprehensively consider multidimensional consequences including such as death, property loss, and environmental pollution. Instead, it treats these as separate issues. As these disparate outcomes are closely interrelated and frequently stem from similar underlying causal events, modeling them in isolation prevents information sharing between tasks. This structural constraint may result in reduced predictive accuracy and less reliable understanding of an incident's comprehensive impact [17-19].

In recent years, interpretability techniques such as LIME and SHAP have been employed to analyse the decision logic of complex predictive models [1, 20]. These methods can estimate feature importance based on input–output relationships without requiring access to the internal model structure [21]. However, most of these approaches provide only post hoc explanations and lack explicit causal perspectives or mechanistic foundations, often leading to instability and low-fidelity interpretations. To overcome these limitations, an emerging line of research has introduced causal inference theory, conceptualizing accident evolution as a system of interacting variables within a structural causal model (SCM) [22]. By employing tools such as causal diagrams, counterfactual reasoning, and mediation analysis, these approaches aim to uncover the underlying mechanisms of accident progression [23-25]. This causal modeling perspective has begun to gain traction in high-risk domains such as transportation and aviation, demonstrating stronger explanatory power and greater potential for intervention-oriented safety strategies.

In summary, despite progress, current risk modeling is constrained by a reliance on spurious correlations and unstable post-hoc explanations, which typically capture only static, average effects. This highlights the need for a paradigm shift from correlation-based black-box models to frameworks with inherent structural interpretability. We therefore propose a causal framework designed to elucidate how latent risk factors evolve into diverse consequences through their structural interactions. Our goal is to provide robust, causal-based decision support for proactive risk management in the maritime sector.

2.2 Dealing with imbalanced data

In machine learning models for accident severity classification, class imbalance

remains a persistent challenge. Existing methods to address this issue can be broadly categorized into undersampling and oversampling [26]. Oversampling avoids the data loss inherent in undersampling, which can remove informative samples and impair the model's generalization ability.

Among oversampling techniques, the synthetic minority oversampling technique (SMOTE) proposed by Chawla et al. [27] has been extensively applied in accident and safety-related studies. Traditional supervised learning algorithms—such as SVM, random forests, and XGBoost—are frequently combined with SMOTE for severity prediction. These methods are relatively easy to implement; however, SMOTE generates synthetic samples via linear interpolation, considering only spatial proximity. As a result, it neglects latent feature structures and may introduce boundary ambiguity or intra-class distortion. This issue becomes particularly critical when minority-class samples lie near the decision boundary, where SMOTE is prone to overfitting and misclassification.

To address these shortcomings, numerous SMOTE variants have been proposed, such as SL-SMOTE [28], LN-SMOTE [29], and SMOTE-IPF [30]. These approaches incorporate boundary information, density estimation, or noise identification to improve the quality of synthetic samples. Yi et al. [31] proposed ASN-SMOTE, which selects qualified minority class instances through k-nearest neighbors (KNN) and uses them to generate new minority instances.

Recently, more advanced generative models have been introduced to enhance minority-class diversity. For instance, Li et al. [32]applied variational autoencoders (VAEs) to learn continuous latent-space distributions and thereby overcome the limitations of traditional sampling. Similarly, Qiao et al. [2] employed generative adversarial networks (GANs) to produce realistic minority-class samples. Despite their success in image generation, GANs often suffer from unstable training when applied to tabular data with discrete features, due to their reliance on Kullback–Leibler divergence. In contrast, diffusion probabilistic models (e.g., DDPMs) simulate a forward–reverse diffusion process in the feature space, enabling the generation of diverse, distribution-consistent samples and demonstrating superior performance in imbalanced data scenarios [33].

Nevertheless, a critical limitation persists: most sampling approaches train generative and classification models separately, with little interaction between them. This separation prevents the generated samples from being optimized with respect to

classifier weaknesses, limiting their practical effectiveness. Future research should therefore focus on collaborative or joint training mechanisms in which generative models adaptively refine sample distributions to complement the classifier's decision boundaries. Such integration can not only strengthen minority-class recognition but also substantially enhance model robustness and predictive performance in real-world imbalanced settings.

3. Methodology

3.1 Overview of the proposed methodology

The methodological framework proposed in this study is shown in Fig. 1. Its core process encompasses four main phases: data collection and statistics (A), data synthesis (B), core model construction (C), and model evaluation and interpretation (D).

Accident reports were collected from a global maritime database, and basic accident information was extracted. Statistical analysis (as shown in Fig. 1-A) revealed a significant class imbalance in the data regarding accident severity, with the minority class (e.g., severe accidents) having fewer samples than the majority class. This severely impacts the model's learning performance and generalization capabilities. To address this data imbalance, a data synthesis module (as shown in Fig. 1-B) was designed. This module compares and integrates various advanced generative models (such as SMOTE, CTGAN [34], and TabDDPM) to generate high-quality synthetic data. Specifically, a co-training mechanism for the classifier and generator (as shown in Fig. 1-C) was designed. Information provided by the classifier guides the generation process, ensuring that the synthetic data both maintains the distributional characteristics of real data and effectively supplements scarce samples. The core of this research is an innovative deep causal decoupling network (as shown in Fig. 1-C). This model dynamically explores and identifies key heterogeneous mediating patterns from each accident sample through a cross-attention mechanism. Subsequently, through a decoupled framework integrating residual structure, orthogonal constraints, and gradient reversal adversarial training, it enforces the separation of the original information flow into independent direct causal, indirect causal, and non-causal features. After obtaining the decoupled causal representation, a dynamic fusion strategy is used to perform task predictions (accident severity as the primary task, and casualties, property losses, and environmental pollution as auxiliary tasks), resulting in more robust and informed accident classification decisions. This significantly enhances the

model's predictive robustness and inherent interpretability. To comprehensively evaluate the model's performance and interpretability, we not only conducted comprehensive performance validation on multiple prediction tasks but also designed specialized causal analysis experiments (as shown in Fig. 1-D). Using methods such as counterfactual inference, we quantified the impact of different causal paths on the final accident classification, providing deeper and more reliable insights for maritime safety decision-making.

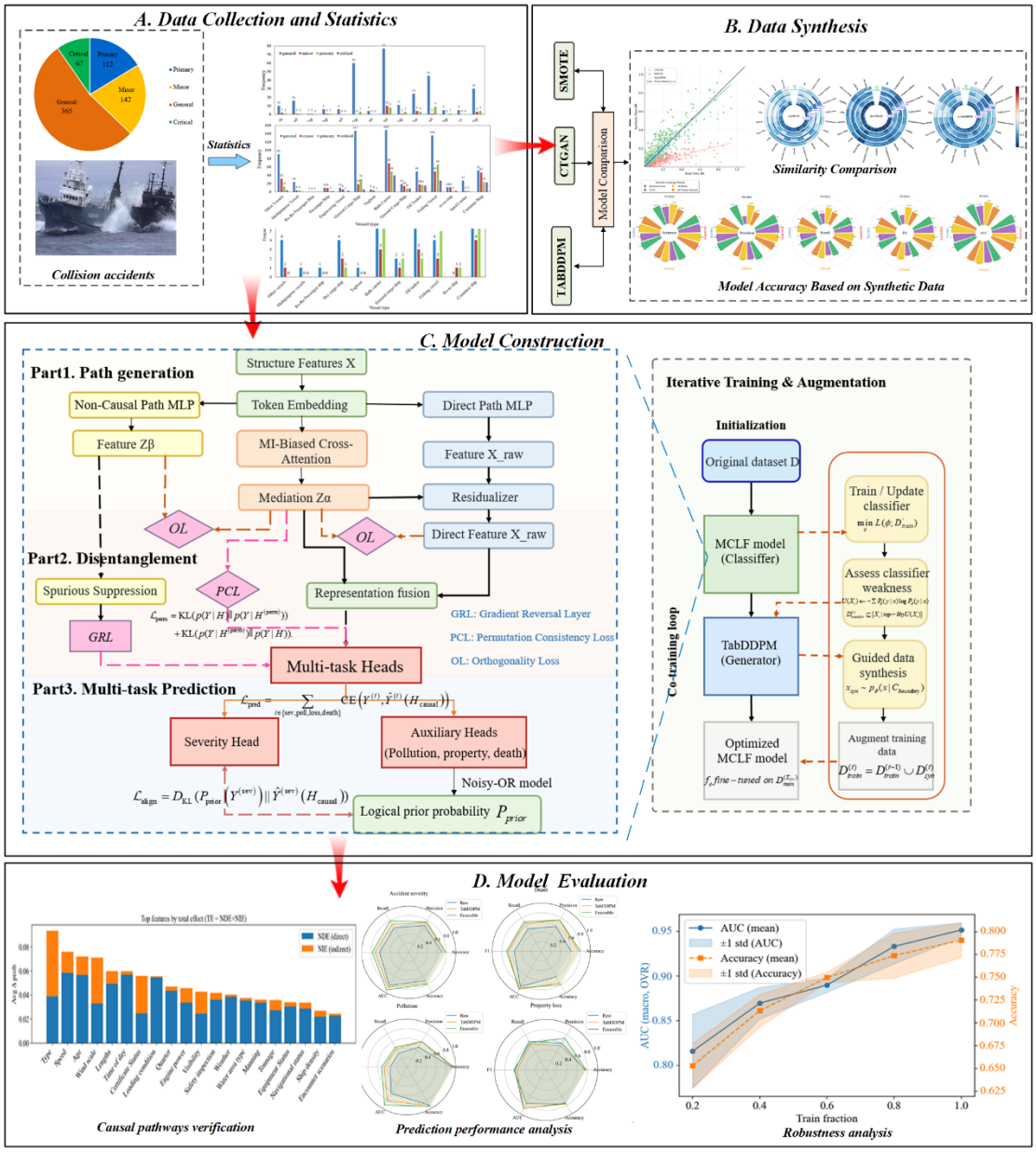


Fig. 1. General process of the present study.

3.2 Data source

Maritime accident reports published were collected from the International Maritime Organization (IMO) as well as national maritime authorities, including the China

Maritime Safety Administration (China MSA), the United States National Transportation Safety Board (NTSB), and various European maritime agencies. Each selected report provides detailed information on the accident, including the time and location of occurrence, vessel condition, weather conditions, and accident characteristics. To ensure data completeness, a total of 686 accident reports are used to demonstrate the effectiveness and feasibility of the proposed methodology.

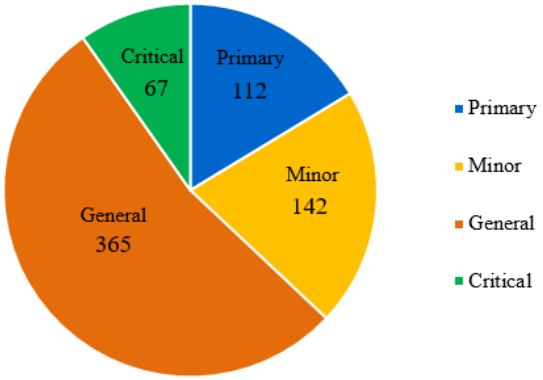


Fig. 2. Statistics on the number of accidents by severity level.

Currently, individual countries have also developed their own classification standards. We adapted and simplified these classification systems into four categories—minor, general, primary, and critical accidents—to facilitate model development and evaluation. The classification excludes catastrophic accidents due to their rarity; in the dataset, only one such case of a catastrophic collision was identified. The distribution of the 686 ship collision accident reports used in this study is as follows: 142 minor accidents (20.7%), 365 general accidents (53.2%), 112 primary accidents (16.3%), and 67 critical accidents (9.8%).

However, this dataset has an obvious class imbalance problem: the number of samples for primary and critical accidents and large accidents is less than that of other accidents. This distribution will cause the model to be biased towards the dominant category during training, thereby affecting the ability to identify high-severity accidents. To solve this problem, a synthetic sample enhancement mechanism is introduced.

3.3 Causal perspective

In this work, we revisit the task of maritime accident severity prediction through the lens of causal learning. Traditional approaches typically model the relationship between structured input features and severity labels via supervised learning, often overlooking the underlying causal mechanisms that govern such relationships.

However, in complex socio-technical systems like maritime operations, ignoring causal structure can result in spurious correlations, poor generalization, and lack of interpretability. We do not assume unsafe factors are the known inputs. Instead, we use the observable antecedent conditions (X) to infer the probability distribution of potential unsafe behaviors. This approach is not only more aligned with the reality of proactive risk management but also enhances interpretability by revealing how situational factors lead to risk, thus paving the way for targeted interventions.

To address these challenges, we formulate a Structural Causal Model (SCM) [35] that explicitly captures the causal dependencies among key variables. This model enables us to reason about interventions and disentangle genuine causal effects from statistical associations.

Let the SCM be defined over the following variables:

- $X \in R^d$: Structured input features (e.g., vessel attributes, environmental factors, navigation features)
- $Z \in R^m$: Intermediate unsafe factors features (e.g., untimely collision avoidance)
- $Y \in R$: Target variable representing accident severity
- U: Spurious features

The corresponding causal diagram is:

- $X \rightarrow Z \rightarrow Y$: This path captures the indirect causal influence of structural features X on severity Y through intermediary factor Z. Z reflects the most likely unsafe factors under current navigation conditions. This shift from diagnostic causal models to predictive causal models is crucial for maritime risk management. For instance, adverse environmental conditions(X) may increase the likelihood of navigation errors (Z), which subsequently lead to more severe accidents.
- $X \leftarrow U \rightarrow Y$: The spurious factors U may introduce superficial correlations between variables X and outcomes Y, but do not reflect any actionable or stable mechanism
- $X \rightarrow Y$: Represents the direct influence of structured features on the outcome, independent of Z.

3.4 Analysis and objectives

Many current data-driven approaches, especially neural network-based models, suffer from a lack of interpretability and a tendency to exploit spurious patterns in the

training data, thereby compromising both their trustworthiness and generalization capabilities.

This problem is particularly critical in multi-target risk scenarios, where structured data (e.g., vessel behavior, environment, vessel properties) interact with latent unsafe factors (e.g., untimely avoidance, bad weather, improper collision avoidance) to affect multiple outcomes (e.g., casualties, property losses, pollution). In such settings, it is not sufficient merely to achieve high predictive accuracy; the following questions must also be answered:

- Which features contribute to high-risk outcomes?
- Are the effects direct or mediated through latent unsafe conditions?
- How robust are these findings under changing environments or interventions?

To address these challenges, we propose a causal information disentanglement framework that explicitly decouples and controls causal and non-causal information. Our framework is built upon the central hypothesis that accident severity outcomes are not merely correlated with observed features, but are mediated through a set of latent unsafe factors, a concept rooted in classical causal mediation analysis[36]. To operationalize this hypothesis and ensure effective disentanglement, our model is designed to satisfy three key properties derived from information theory. This approach is inspired by the Information Bottleneck principle [37], which seeks a compressed representation that is maximally informative about a target. However, to achieve causal disentanglement, we evolve this principle: instead of applying a general compression, we enforce a structured regularization. This allows our model to explicitly isolate the information flowing through the causal factors while actively suppressing spurious correlations captured by non-causal representations:

Based on this insight, three information theory properties must be met:

(Purpose A): $$\max I\left(\left(X_{\alpha}, Z_{\alpha}\right); Y\right) \quad (1)$$

Where X_{α} denotes the direct causal information from the input, and Z_{α} represents the information transmitted through latent mediators. This ensures that the representation is information-complete with respect to predicting accident severity, leaving out no essential causal signal.

(Purpose B): $$I\left(Z_{\beta}; Y\right) \approx 0 \quad (2)$$

Where Z_β denotes the latent subspace that captures only non-causal correlations with the outcome. By suppressing these dependencies, the model prevents reliance on misleading shortcuts.

Different latent subspaces must be mutually independent to avoid redundancy and information leakage.

(Purpose C): $$X_\alpha \perp Z_\alpha, Z_\alpha \perp Z_\beta \tag{3}$$

In accordance with the principles outlined above, we construct an integrated learning objective function $\mathcal{L}$ to co-optimize three goals: (1) enhancing the predictive power of causal pathways, (2) suppressing spurious correlations, and (3) promoting representation disentanglement. The function is formulated as:

$$\mathcal{L} = -I\left(\left(X_\alpha, Z_\alpha\right); Y\right) + \beta \cdot I\left(Z_\beta; Y\right) + \gamma \cdot \left(I\left(X_\alpha; Z_\alpha\right) + I\left(Z_\alpha; Z_\beta\right)\right) \tag{4}$$

Where β and γ are hyperparameters that balance the different objectives.

This disentanglement guarantees that causal (direct and mediated) and non-causal (spurious) information are explicitly separated. As a result, the output can be additively decomposed as,

$$\hat{Y} = g_d\left(X_\alpha\right) + g_m\left(Z_\alpha\right), \tag{5}$$

Thereby making the total effect explicitly interpretable as the sum of a direct effect $g_d\left(X_\alpha\right)$ and a mediated effect $g_m\left(Z_\alpha\right)$. This decomposition provides a principled foundation for both predictive accuracy and causal interpretability.

3.5 Disentangling causal pathways for interpretable prediction

3.5.1 MI-Biased cross-attention

Let the input be a set of categorical features $X = \left(X_1, \ldots, X_j\right)$, where each feature X_j is discrete, $X_j \in \left\{0, \ldots, C_j - 1\right\}$. For supervision, we are given a set of labels for unsafe factors, $Z = \left(Z_1, \ldots, Z_K\right)$ with $Z_k \in \left\{0,1\right\}$, and a set of downstream multi-task labels $Y = \left\{Y^{(\text{sev})}, Y^{(\text{poll})}, Y^{(\text{loss})}, Y^{(\text{death})}\right\}$.

Each feature X_j is first mapped to a dense vector token $T_j \in \mathbb{R}^d$ via an embedding function e_j:

$$T_j = e_j\left(X_j\right), T = \left[T_1, \ldots, T_j\right] \in \mathbb{R}^{F \times d}. \tag{6}$$

These tokens are then linearly projected to generate the key (K) and value (V) matrices for the attention mechanism:

$$K = TW_K, V = TW_V. \tag{7}$$

Where $W_K, W_V \in \mathbb{R}^{d\times d}$ are the learnable weights.

To identify each mediating factor Z_k, we introduce a set of learnable query vectors $Q = [q_1, \ldots, q_K]^\top \in \mathbb{R}^{K\times d}$. A standard scaled dot-product attention score between query q_k and key K_j is computed as:

$$\text{base}_{k,j} = \frac{\langle q_k, K_j \rangle}{\sqrt{d}}. \tag{8}$$

Crucially, we inject a strong inductive bias by incorporating a pre-computed feature-mediator Mutual Information (MI) matrix $M \in \mathbb{R}^{j\times K}$ as a static prior. Each element $M_{j,k}$ is proportional to the mutual information $I(X_j; Z_k)$, estimated offline from the dataset using a consistent discrete MI estimator and normalized per column. The final attention logits are formed by adding this bias:

$$\text{logits}_{k,j} = \text{base}_{k,j} + \lambda M_{j,k}, \tag{9}$$

Where $\lambda > 0$ is a hyperparameter controlling the strength of the MI prior. This bias guides the attention mechanism to focus on features that are statistically most relevant to a given mediator from the outset of training. The attention weights $\pi_{k,j}$ are then obtained via softmax over the feature dimension:

$$\pi_{k,j} = \frac{\exp(\text{logits}_{k,j})}{\sum_{j'=1}^{j} \exp(\text{logits}_{k,j'})}. \tag{10}$$

The contextual embedding for each mediator, $H_k^{(Z)}$ is computed by attending to the value vectors V using the MI-biased weights $\pi_{k,\cdot}$:

$$H_k^{(Z)} = \sum_{j=1}^{F} \pi_{k,j} V_j \in \mathbb{R}^d. \tag{11}$$

The activation probability of each mediator, $p_k = \mathbb{P}(Z_k = 1 \mid X)$, is predicted using a shared linear head (ω, b) on top of its embedding:

$$\ell_k = \omega^\top H_k^{(Z)} + b, p_k = \sigma(\ell_k). \tag{12}$$

We then compute an aggregated mediator representation z_α by taking a weighted average of their embeddings $H_k^{(Z)}$, where α_k the weights are the normalized activation probabilities:

$$\alpha_k = \frac{p_k}{\sum_{k'} p_{k'} + \varepsilon} \tag{13}$$

$$Z_\alpha = \sum_{k=1}^{K} \alpha_k H_k^{(Z)} \in \mathbb{R}^d. \tag{14}$$

3.5.2 Decoupling pathways

We train a residualizer R_ϕ to reconstruct the portion of X that is predictable from Z_α:

$$\hat{X} = R_\phi(Z_\alpha), X_\alpha = X - \text{stopgrad}(\hat{X}). \tag{15}$$

The residualization loss is:

$$\mathcal{L}_{\text{resid}} = \| X - \hat{X} \|_2^2 . \tag{16}$$

Thus, by construction, X_α represents the variation in X that is orthogonal to (or decorrelated from) the mediator-explained representation Z_α. What remains is the direct effect of external conditions on Y independent of mediators. The stopgrad operator prevents leakage of information from the residualizer R_ϕ back into X_α.

To ensure that Z_α robustly captures mediating information and is not confounded by direct predictive signals, we employ a suite of three complementary regularization techniques.

(a) Orthogonal Decorrelation ($L_\perp$): To enforce pathway disentanglement, we penalize the linear correlation between the direct (X_α) and mediating (Z_α) representations. We define the batch-wise covariance matrix $C(u,v) = \frac{1}{B-1}\tilde{u}^\top \tilde{v}$, where $\tilde{u}$ is the centered version of u. The loss is:

$$\mathcal{L}_\perp = \lambda_{xa} \| C(x_\alpha, z_\alpha) \|_j^2 + \lambda_\beta \| C(z_\beta, z_\alpha) \|_j^2 . \tag{17}$$

This loss acts as a proxy for the disentanglement objective (L_{disen}), driving the representations into approximately orthogonal subspaces. (Purpose C)

(b) Adversarial Debiasing (Dual GRL): To actively remove spurious correlations, we employ two adversarial objectives using the Gradient Reversal Layer (GRL).

1. Spurious path suppression ($Z_\beta \to Y$): A prediction head is trained via GRL to predict the main task $Y^{(\text{sev})}$ from the bypass representation Z_β. The main model is trained to *maximize* this head's loss ($\mathcal{L}_{\text{adv}}$), thereby forcing Z_β to be uninformative about Y, analogous to minimizing $I\left(Z_\beta; Y\right)$. (Purpose B)

2. Anti-mediation for direct path ($X_\alpha \to Z$): Similarly, another head is trained via GRL to predict the mediators Z from the direct path representation X_α. By maximizing this loss ($\mathcal{L}_{X\to Z}$), we ensure that information predictive of Z is routed through the cross-attention mechanism (Z_α) rather than leaking into the direct path X_α.

(c) Permutation consistency for interventional sensitivity (L_{perm}): To enforce the causal necessity of the mediator, we introduce a permutation consistency loss. This loss is designed to make the model's predictions sensitive to interventions on the mediator representation Z_α. Specifically, we shuffle Z_α along the batch dimension to create a counterfactual representation, denoted as Z_α^{perm}. Let $H = \phi\left(\left[X_\alpha \,\|\, Z_\alpha\right]\right)$ be the combined representation before the final prediction layer, and $H^{\text{perm}} = \phi\left(\left[X_\alpha \,\|\, Z_\alpha^{\text{perm}}\right]\right)$ be its counterfactual counterpart. The model is penalized if its prediction *fails* to change significantly after this intervention. This is achieved by maximizing the symmetric KL-divergence between the original and counterfactual predictive distributions. Equivalently, we minimize the following loss:

$$\mathcal{L}_{\text{perm}} = \text{KL}(p(Y \mid H) \| \, p(Y \mid H^{(\text{perm})})) + \text{KL}(p(Y \mid H^{(\text{perm})}) \| \, p(Y \mid H)). \tag{18}$$

Minimizing this objective encourages the model to be highly sensitive to the content of Z_α, thereby reinforcing its role as a necessary component in the causal chain from $X \to Z \to Y$.

3.5.3 Multi-task prediction on causal representations

Following disentanglement, we fuse the direct path (X_α) and mediator path (Z_α) representations into a unified causal representation, $H_{\text{causal}} = \text{concat}\left(X_\alpha, Z_\alpha\right)$. This representation serves as the shared input for a multi-task learning framework designed

to jointly predict the main task (severity, $Y^{(\text{sev})}$) and three logically correlated auxiliary tasks (pollutant, loss, and casualty events).

All tasks are optimized via a joint prediction loss, $\mathcal{L}_{\text{pred}}$ defined as the sum of their Cross-Entropy losses (Purpose A):

$$\mathcal{L}_{\text{pred}} = \sum_{t\in\{\text{sev,poll,loss,death}\}} \text{CE}\left(Y^{(t)}, \hat{Y}^{(t)}\left(H_{\text{causal}}\right)\right) \tag{19}$$

Minimizing this loss forces the upstream encoders to learn high-quality causal representations (X_α, Z_α) that are robust enough to support all correlated tasks.

To strengthen the logical consistency of prediction, a knowledge fusion mechanism is introduced to construct a logical prior for "severity" based on the marine accident classification standard, thereby constraining the prediction behavior of the model. It constructs a logical prior for severity $P_{\text{prior}}\left(Y^{(\text{sev})}=1\right)$ by fusing evidence from the auxiliary tasks using a Noisy-OR model:

$$P_{\text{prior}}\left(Y^{(\text{sev})}=1\right)=1-\prod_{t\in\{\text{poll,loss,death}\}}(1-\hat{P}^{(t)}) \tag{20}$$

Where $\hat{P}^{(t)} = P(Y_t = 1 | H_{causal})$ is the predicted probability for each auxiliary task t. An alignment loss $\mathcal{L}_{\text{align}}$ then regularizes the main prediction by minimizing the KL divergence between the model's output and this logical prior:

$$\mathcal{L}_{\text{align}} = D_{\text{KL}}(P_{\text{prior}}\left(Y^{(\text{sev})}\right) || \hat{Y}^{(\text{sev})}\left(H_{\text{causal}}\right)) \tag{21}$$

This ensures the model's predictions are not only data-driven but also logically coherent.

3.5.4 Overall Objective Function

The model is trained end-to-end by minimizing a composite objective function $\mathcal{L}_{total}$, which is structured to balance two fundamental goals: predictive accuracy and causal validity. This is achieved by combining a core supervised loss with a causal regularization term:

$$\mathcal{L}_{total} = \mathcal{L}_{supervised} + \mathcal{L}_{causal\_reg} \tag{22}$$

$\mathcal{L}_{supervised}$ is the core supervised loss that anchors the model to the ground-truth data. It consists of the loss for mediator prediction and the loss for the downstream tasks. $\mathcal{L}_{causal\_reg}$ is the causal regularization term, which imposes critical structural constraints

on the learned representations to ensure their causal soundness and interpretability. This composite objective function serves as the tractable proxy for the ideal but intractable information-theoretic goals of causal representation learning. By jointly optimizing for prediction accuracy and these structural constraints, our model learns to identify and represent unsafe factors as interpretable mediators of accident severity.

3.6 Dynamic augmentation for multi-task imbalance

In many practical machine learning applications, the problem of class imbalance is widespread. This imbalance in data distribution often causes traditional classifiers to favor the majority class, which significantly reduces the detection performance of the model on the minority class, especially in tasks that require extremely high recognition of the minority class, such as medical diagnosis and traffic accident prediction. To alleviate this problem, researchers have proposed a variety of oversampling methods, among which SMOTE and its variants (such as Borderline-SMOTE [38], ADASYN [31], MLSMOTE [39]) have been widely used. However, they suffer from two key limitations in our context: (1) They are typically static, operating as a pre-processing step without adapting to the classifier's learning process. (2) They are often designed for single-task scenarios and struggle to effectively model the complex, high-dimensional distributions of multi-task label combinations.

To overcome the above shortcomings, this paper introduces a dynamic generation framework based on the tabular diffusion model TabDDPM [33]. It is a generative model for tabular data using a diffusion-based approach. It models both numerical and categorical data types through a combination of Gaussian diffusion for continuous features and multinomial diffusion for categorical features. We use it in conjunction with the classifier, as shown in Algorithm 1.

Algorithm 1: Dynamic Augmentation for Multi-Task Imbalance

function MTIA($D, \varepsilon_\theta, N_{target}, w, f_\phi, T_{iter}, E, k, N_{batch}$):

// Phase 1: Initialization

Build frequency map $N(C)$ from original dataset D;

Compute required samples per combination:
$N_{\text{gen}}(C) \leftarrow \max(0, N_{\text{target}} - N(C))$;

Construct initial generation task list $\mathcal{L}_{\text{gen}}^{\text{init}}$ based on $N_{\text{gen}}(C)$;

Generate initial synthetic dataset $D_{\text{syn}}^{(0)}$ using TabDDPM;

// Update rule is applied during generation

$\widehat{\varepsilon}(x_t, C, t) \leftarrow \varepsilon_\theta(x_t, , t) + w\big(\varepsilon_\theta(x_t, C, t) - \varepsilon_\theta(x_t, , t)\big)$;

$D_{\text{train}}^{(0)} \leftarrow D \cup D_{\text{syn}}^{(0)}$;

// Phase 2: Iterative Training and Augmentation

for $t \leftarrow 1$ **to** T_{iter} **do**

Train predictor f_ϕ on $D_{\text{train}}^{(t-1)}$ for E epochs;

Compute uncertainty for each sample $X_i \in D$:
$U(X_i) \leftarrow -\sum_j p_\phi(y_j|X_i) \log p_\phi(y_j|X_i)$;

Select top-k most uncertain samples to form the boundary set $D_{\text{boundary}}^{(t)}$;

Build a new task list $\mathcal{L}_{\text{gen}}^{(t)}$ for data generation at the boundary;

Generate new synthetic boundary data $D_{\text{syn}}^{(t)}$ via TabDDPM;

Update training set: $D_{\text{train}}^{(t)} \leftarrow D_{\text{train}}^{(t-1)} \cup D_{\text{syn}}^{(t)}$;

// Phase 3: Final Training

Fine-tune predictor f_ϕ on the final augmented dataset $D_{\text{train}}^{(T_{\text{iter}})}$ until convergence;

$\phi^\star \leftarrow \arg\min_\phi \mathcal{L}_{\text{MT}}\big(\phi; D_{\text{train}}^{(T_{\text{iter}})}\big)$;

return $\phi^\star$ *// Return the optimized predictor parameters*

3.7 Evaluation metrics

In multi-classification tasks, the confusion matrix can be used to intuitively represent the predictions of the classifier on each category.

Assume there are K categories in total. The element N in the *i-th* row and *j-th* column of the confusion matrix represents the number of samples whose true category is the *i-th* category and is predicted to be the *j-th* category. The following performance evaluation metrics can be defined [1]:

Accuracy is the ratio of correctly classified samples to the total number of samples:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

Precision is the ratio of correctly predicted observations in the particular class to all predicted observations in the same class, which is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{24}$$

Recall is the ratio of correctly predicted observations under a given class to all reference observations in that class, which is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{25}$$

F1-score is an indicator calculated from recall and precision, which is expressed as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{26}$$

AUC provides a threshold-independent measure of the model's overall performance in discriminating between positive and negative classes.

$$ROCAUC = \int_0^1 TruePositiveRated(FalsePositiveRate) \tag{27}$$

4. Results and discussion

To validate the effectiveness and interpretability of our proposed approach, we designed a series of comprehensive experiments aimed at answering the following key research questions derived from our objectives.

RQ1: Impact of Data Synthesis on Predictive Performance. Given the inherent class imbalance in maritime accident data, which data synthesis technique (e.g., SMOTE, CTGAN, TABDDPM) most effectively improves the performance of our predictive model?

RQ2: Predictive Performance. Does our causal model achieve superior or at least competitive performance in predicting maritime accident severity when compared against a suite of strong non-causal baseline models?

RQ3: Validation of the Causal Mechanism. Can we empirically verify that the learned latent representation, Z, functions as a genuine causal mediator in the X→Z→Y pathway, rather than being a mere statistical correlate?

RQ4: Causal Effect Decomposition and Interpretation. How does our model quantify and disentangle the direct and indirect effects of critical risk factors (e.g., vessel type, time of day), and what novel insights into accident causation can be derived from this decomposition?

4.1 Descriptive analysis of accident factors

In the present study, pre-accident factors can be divided into three categories: ship-related characteristics (i.e., ship type, tonnage, length, width, age, main engine power,

loading, registration survey status, and safety inspection status), environment-related characteristics (i.e., water type, time, visibility, wind speed, season, and channel conditions), and navigation-related characteristics (i.e., encounter scenario, speed, navigation status, and density). In addition, based on previous studies, the classification of the Maritime Accident Investigation Bureau, which is widely used in the industry, is divided into states[40]. The state classification of each characteristic is given in Table 1 (Appendix).

In the present study, we selected four key outcome variables—property losses, death severity, pollution and accident severity—as the focal dimensions of accident impact. To explore the interdependencies among the selected outcomes and to guide modeling design, we conducted a correlation and significance analysis using both Pearson correlation coefficients, maximal information coefficients (MIC), and p-values, as shown in Fig. 4.

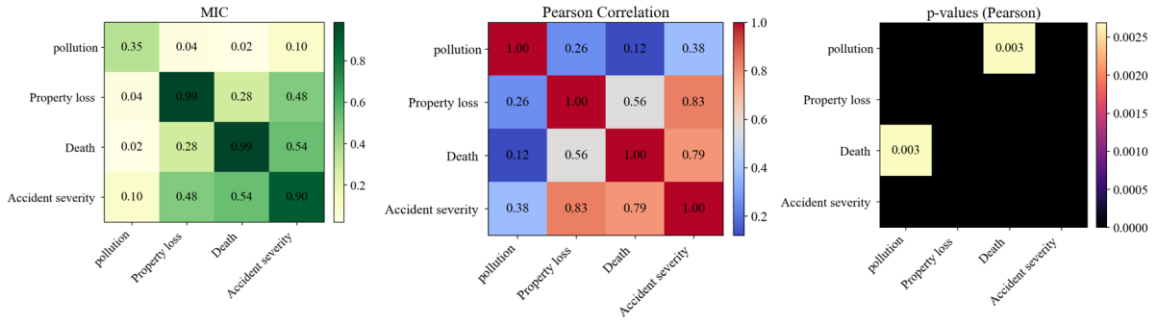


Fig. 4. Correlation analysis of consequence variables.

Based on the statistical analysis presented in Fig. 4, we can observe significant interdependencies among the accident-related variables, which provides a clear rationale for our modeling choice. The heatmaps for both Maximal Information Coefficient (MIC) and Pearson Correlation reveal a strong, positive, and statistically significant relationship between 'Accident severity', 'Property loss', and 'Death'. Specifically, 'Accident severity' shows a very high linear correlation with 'Property loss' ($r = 0.90$, $p < 0.001$) and a strong linear correlation with 'Death' ($r = 0.52$, $p < 0.001$). Furthermore, 'Property loss' and 'Death' are themselves moderately correlated ($r = 0.46$, $p < 0.001$). This indicates that these three variables do not operate independently; rather, they are deeply intertwined aspects of a single event's outcome. Pollution has a weak linear correlation with other variables, but it is not an insignificant noise variable. In fact, the accident severity classification standard itself is defined by three indicators: pollution, property damage, and fatalities. Therefore, pollution remains indispensable when characterizing the overall impact of an accident. Because "accident severity" is a

comprehensive level defined by pollution, property damage, and fatalities, it is suitable as an auxiliary prediction task to help the model learn the overall severity pattern of the accident. Multi-task learning provides a principled approach to leverage these interdependencies: it enables shared representation learning driven by severity, while maintaining task-specific heads for deaths and property losses to account for their distinct variance components. This design enhances generalization by encouraging cross-task regularization and captures both common and task-unique signals within a unified modeling framework.

Regarding the factors and classifications that affect the severity of maritime accidents, we refer to previous studies [41]. By optimizing each report, information about relevant factors can be obtained, and finally a data set for analysis is formed. Table 3 (Appendix) lists the influencing factors selected for analysis and the category description of each factor, and Fig. 5 shows the distribution of the unsafe factors among the accidents.

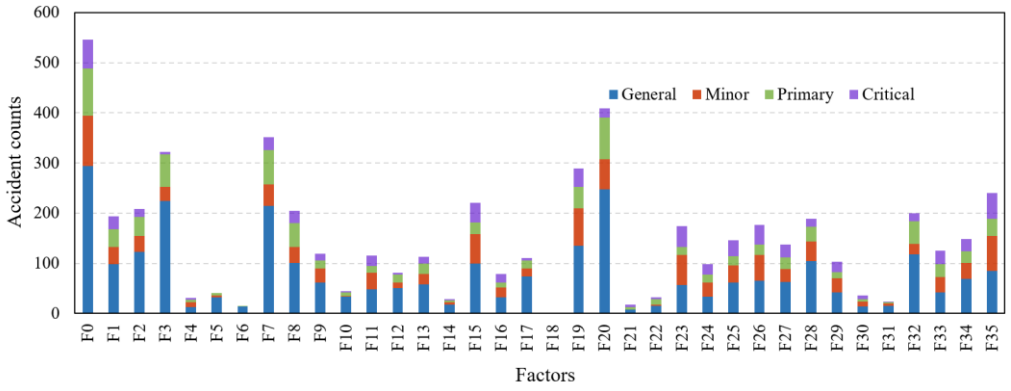


Fig. 5. Distribution of 36 accident factors among the accidents.

4.2 Statistical analysis of the accident

To understand the key factors influencing accident severity in maritime transportation, we conducted a statistical analysis of accident records categorized by vessel type.

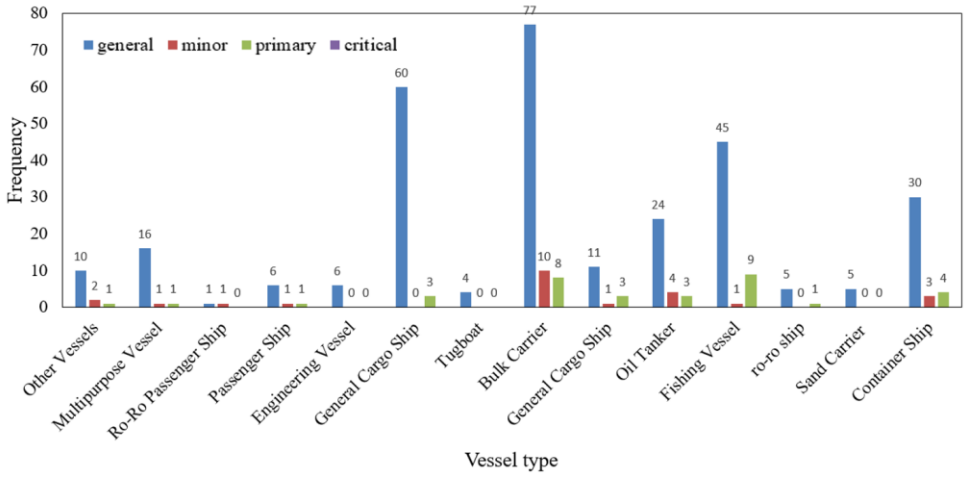


Fig. 6. Relationship between property losses and ship type distribution without casualties.

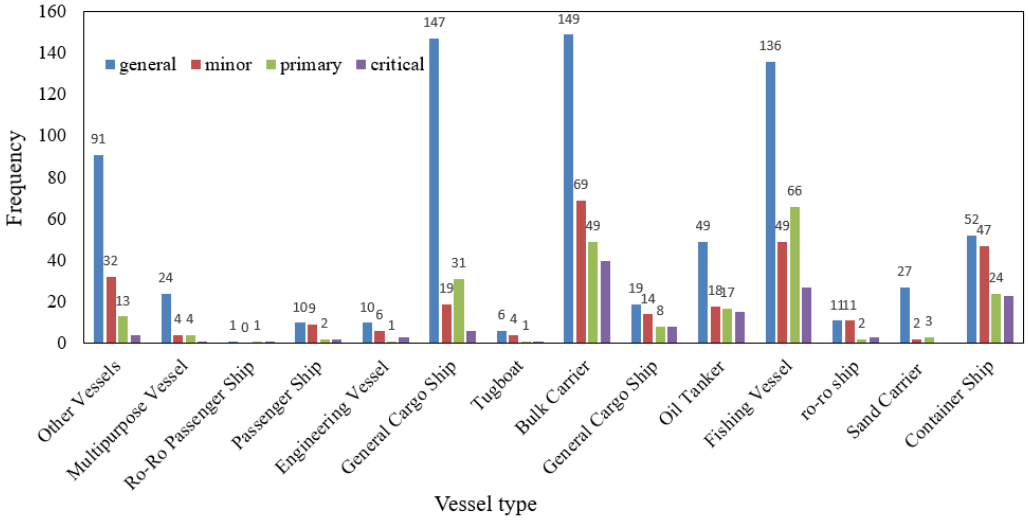


Fig. 7. Relationship between casualties and ship type distribution.

The two figures above reveal distinct patterns in the distribution of vessel types under different accident consequences—specifically in terms of property losses without casualties and personnel injuries or fatalities. These findings provide strong evidence for heterogeneity in accident severity across vessel categories.

In Fig. 6, Bulk Carriers and General Cargo Ships exhibit the highest frequency of accidents with purely economic consequences. Bulk Carriers, in particular, show a notable number of incidents with primary and critical levels of financial loss, suggesting a higher potential for significant damage even in the absence of casualties. Conversely, Tugboats, Ro-Ro Ships, and Sand Carriers show minimal to no involvement in severe property losses, indicating lower risk profiles for these vessel types in such contexts. Fig. 7 shifts focus to personnel outcomes. Bulk Carriers, Fishing Vessels, and Container Ships are prominent contributors to accidents resulting in minor, primary, or critical injuries. Notably, Fishing Vessels stand out with disproportionately high numbers of

severe and critical injury incidents relative to their general accident frequency, suggesting structural or operational vulnerabilities. Container Ships, despite a moderate number of total incidents, frequently result in both property and human losses, indicating high stakes and unpredictability.

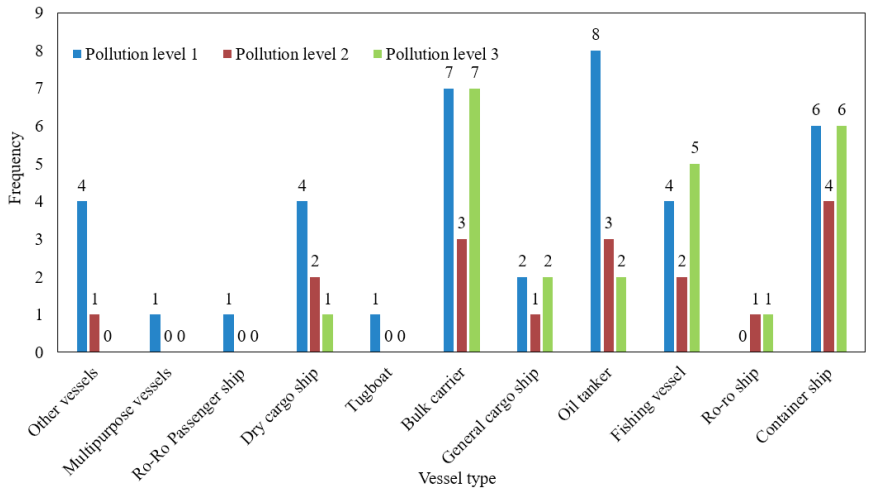


Fig. 8. The frequency distribution of pollution incidents across different ship types.

In Fig. 8, while the majority of incidents are classified at the lowest severity (pollution Level 1), the substantial number of high-severity events (pollution level 3) indicate a persistent threat of major environmental damage. Oil tankers, bulk carriers, and container ships emerge as the most significant contributors to the total number of pollution events. Notably, bulk carriers and container ships display a pronounced risk polarization, with high frequencies at both the lowest (level 1) and highest (level 3) pollution grades. This distribution suggests that while many vessels in these categories operate at a low-risk level, a significant subset poses a severe pollution threat, demanding targeted identification and intervention. Although oil tankers are most frequently involved in level 1 incidents, their continued presence in level 2 and 3 underscores their inherent high-consequence potential. A particularly critical finding is the unexpectedly high frequency of Level 3 incidents involving fishing vessels, which positions them as a major, and potentially underestimated, source of severe pollution. In stark contrast, ship types such as Ro-Ro vessels, multi-purpose ships, and tugs are almost exclusively associated with low-pollution accidents.

4.3 Implementation details

The model proposed in this paper was implemented in the PyTorch framework. The optimal hyperparameter configuration, as detailed in Table 4, was identified through a

systematic tuning process. We employed a random search methodology, which is recognized for its efficiency in exploring high-dimensional hyperparameter spaces.

Table 4 Hyperparameter configuration for model implementation.

Category	Hyperparameter	Value
Model Architecture	Hidden layers	512
	Embedding dimension	128
Training & Optimization	Batch size	256
	Epochs	100
	Learning rate	1e-4
	Weight decay	1e-4
	Dropout rate	0.15
	Orthogonality loss	1e-2
Causal Regularization	MI bias strength	4.0
	Mediation permutation consistency	0.5
	Fusion loss weight	1.0
Adversarial & Decoupling	GRL strength	1.0
	Z adversarial suppression	1.0
	Direct path orthogonality	1e-2
	Residualization	True

4.4 Comparison of synthetic data and original data (RQ1)

Effective synthesis methods must meet dual requirements: (1) Statistical fidelity: the generated data must accurately reflect the complex multivariate distribution of the real data, preserving the marginal distributions of individual features and their complex correlations. (2) Downstream utility: the synthetic data must significantly improve the predictive power and robustness of downstream models when evaluated on a real, unseen test set. Therefore, this experiment aims to determine the most effective data synthesis strategy to create a balanced and high-quality training set, ensuring that all subsequent model comparisons are fair and robust.

We compare three prominent data synthesis techniques:

SMOTE [27]: A classic algorithm that generates new minority class instances by interpolating between existing neighboring instances in the feature space.

CTGAN [42]: A deep learning approach that uses a conditional GAN architecture to learn the data distribution and generate realistic tabular data.

TabDDPM [33]: A generative model based on diffusion processes, which has shown remarkable success in generating high-fidelity data by progressively adding and then learning to remove noise.

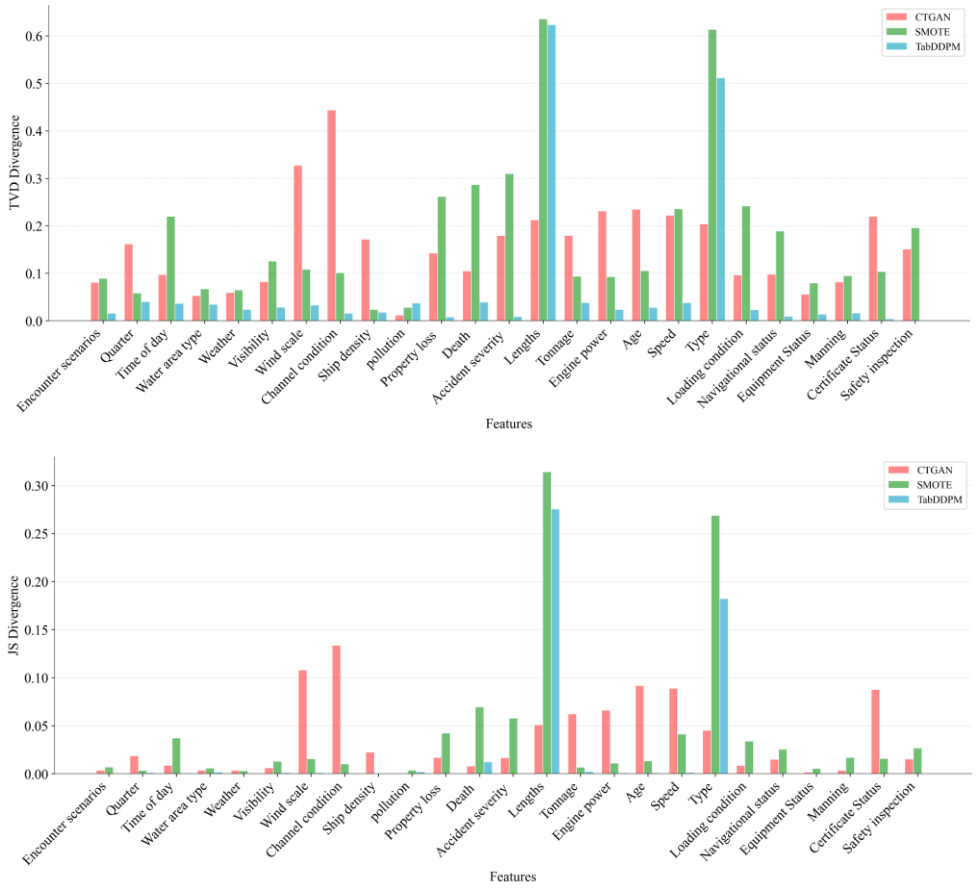


Fig. 9. Comparison of TVD and JS divergence of synthetic data for different models.

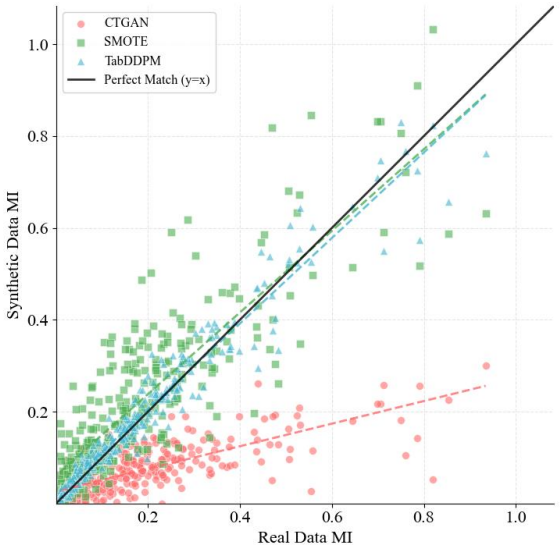


Fig. 10(a). The Scatter Plot of mutual information for real vs. synthetic data from CTGAN, SMOTE, and TabDDPM.

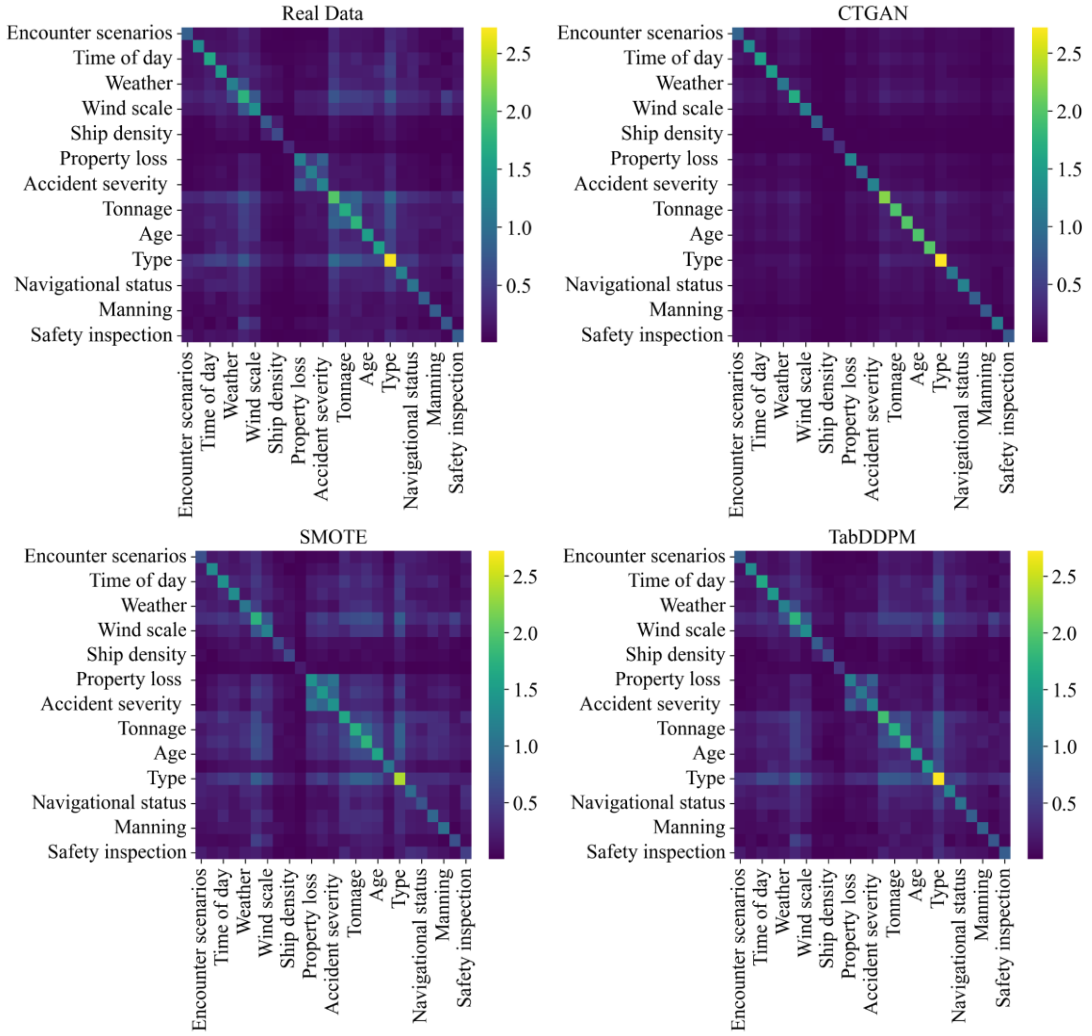


Fig. 10(b). The heatmaps of mutual information for real vs. synthetic data from CTGAN, SMOTE, and TabDDPM.

To comprehensively evaluate the fidelity of the data generation model, we conducted rigorous quantitative and visual analyses along three dimensions: the univariate distribution, the bivariate correlation, and the joint distribution of features with the target variable. Fig. 9 illustrates this evaluation using two statistical metrics—Jensen–Shannon (JS) Divergence and Total Variation Distance (TVD). The results consistently demonstrate the superior ability of the TabDDPM model to reproduce the statistical properties of the original data. Across both sets of histograms, TabDDPM achieves the lowest divergence values in nearly all feature dimensions, with values approaching zero in most cases. In contrast, CTGAN and SMOTE display pronounced deviations from the true distributions.

We further investigated whether the model successfully captured the inherent

correlation structure of the data. By computing and comparing the mutual information between all feature pairs in the real and synthetic datasets, we assessed the model's capacity to learn complex feature dependencies. As shown in Fig. 10(a)(b), TabDDPM's data points are tightly clustered along the diagonal line, and its mutual information heatmap exhibits a high degree of similarity to that of the real data, indicating accurate reproduction of both weak and strong correlations. In comparison, SMOTE preserves a certain positive correlation trend but produces more dispersed and unstable distributions, suggesting weaker consistency in maintaining feature dependencies. CTGAN, on the other hand, performs poorly in reconstructing the complex internal correlation structure present in the original dataset.

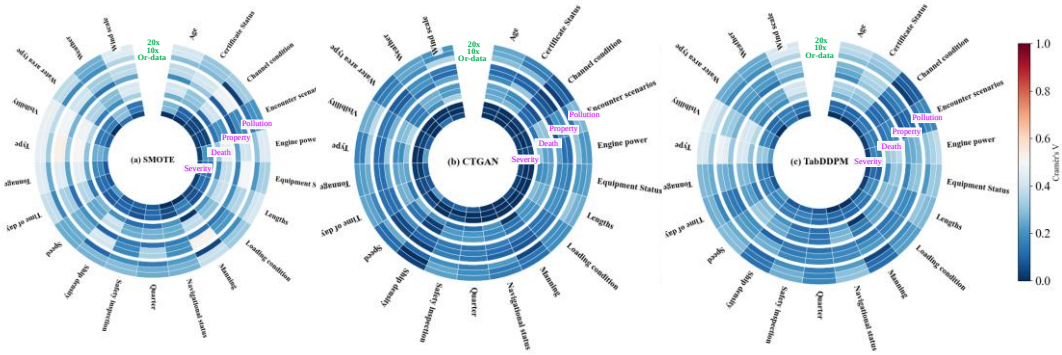


Fig. 11. Comparison of different data augmentation methods on variable correlation.

To assess each model's capacity to preserve the joint distribution between features and target variables, we use the Cramér V correlation coefficient and visualize the results as radial heatmaps (Fig. 11). The color depth, ranging from light to dark blue, represents the correlation strength, with darker shades indicating stronger associations. The TabDDPM model exhibited the highest fidelity. The correlation distribution of its generated data closely matched that of the real dataset, accurately reproducing both strong correlations (e.g., encounter scenario, channel condition) and weak correlations (e.g., age, weather). This indicates TabDDPM's superior ability to capture and replicate the intrinsic dependencies of the data. In comparison, CTGAN preserved the overall distinction between strong and weak correlations but showed noticeable deviations across several features, reflecting moderate fidelity. SMOTE performed worst: its augmented data (middle and outer rings) caused severe degradation of the correlation structure, with many strong correlations in the real data (dark blue) substantially weakened, thereby disrupting key predictive relationships. Overall, these findings confirm that TabDDPM not only generates samples statistically consistent with real

data but also faithfully preserves complex feature–target dependencies, which is essential for constructing reliable predictive models.

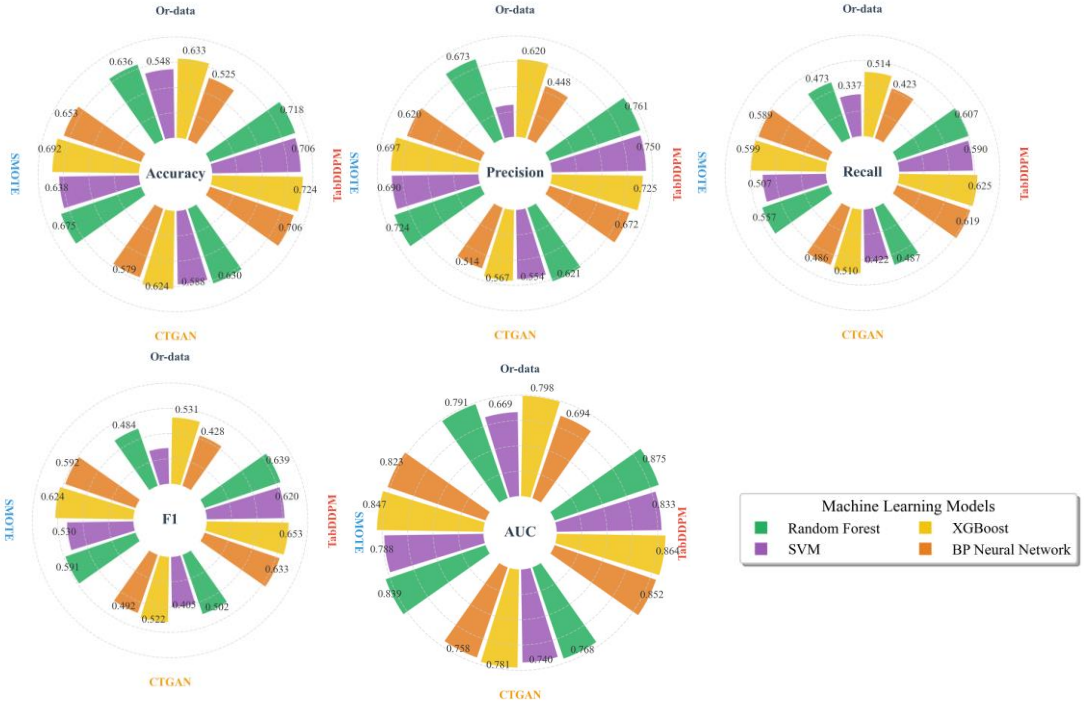


Fig. 12. The prediction effects of different synthetic models.

To assess the practical utility of each synthetic data generation method, we evaluated its impact on downstream modeling tasks. Four widely-used machine learning models—Random Forest (RF), Support Vector Machine (SVM), XGBoost, and a Backpropagation (BP) Neural Network—were trained on the original dataset and augmented datasets from SMOTE, CTGAN, and TabDDPM. As illustrated in Fig. 12, models trained on TabDDPM-generated data consistently achieved the highest performance, substantially outperforming those trained on the original data or data from other synthetic methods. For instance, the RF and XGBoost models trained with TabDDPM data reached AUC scores of 0.875 and 0.864, respectively, marking a significant improvement. This superior performance is attributed to TabDDPM's ability to capture and preserve the complex predictive patterns within the original data, thereby facilitating more robust and effective model training. Conversely, while CTGAN-augmented data generally improved model performance over the original dataset, the gains were markedly smaller than those achieved with TabDDPM. SMOTE demonstrated the most inconsistent results; models trained on its data sometimes underperformed those trained on the original dataset. This suggests its linear interpolation mechanism may introduce noise or obscure decision boundaries, consequently impairing model performance.

In summary, these results underscore that the quality of synthetic data is paramount to its utility in enhancing downstream model performance. The high-fidelity data generated by TabDDPM consistently yields superior training sets that lead to substantial accuracy improvements. This finding validates the efficacy of the TabDDPM approach and highlights the importance of selecting high-fidelity data augmentation strategies, particularly for sparse or imbalanced datasets, to unlock the full potential of predictive models.

4.5 Model prediction performance analysis (RQ2)

Following the identification of TabDDPM as the optimal augmentation method (RQ1), we evaluated the efficacy of our proposed multi-task learning framework against a strong single-task baseline. To ensure a fair comparison, all models were trained on the same TabDDPM-augmented dataset, and their hyperparameters were optimized using a 5-fold cross-validation scheme.

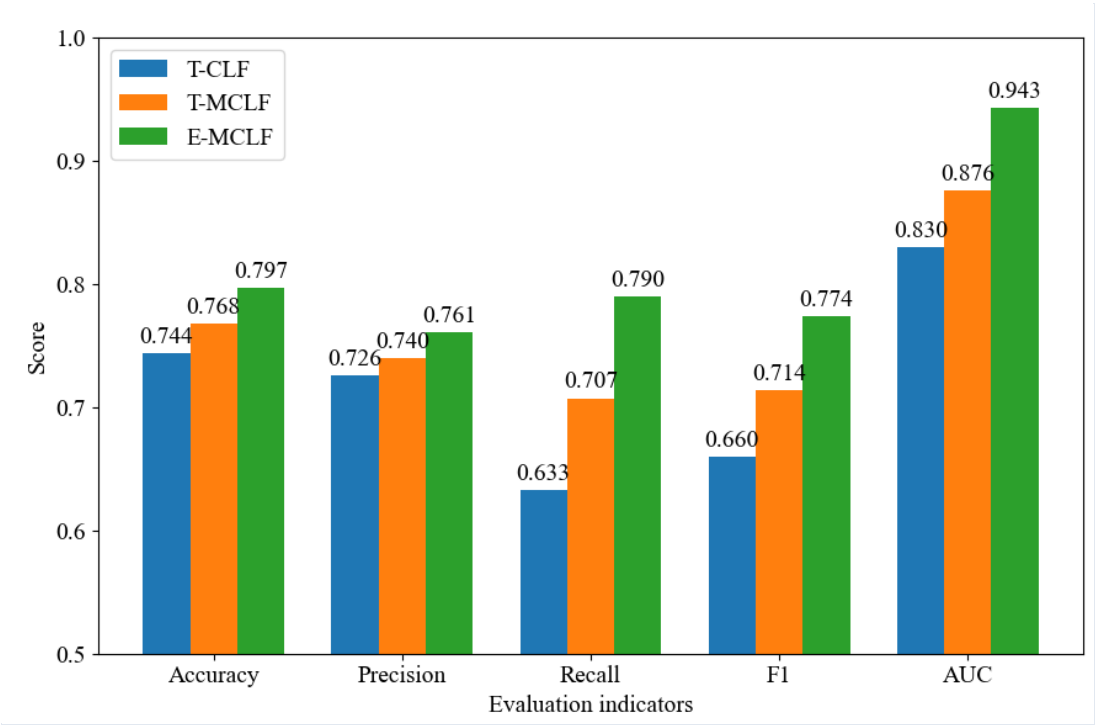


Fig. 13. Comparison of multi-task and single-task prediction performance under different synthesis methods.

The results confirm the superiority of the multi-task approach. Our multi-task model (T-MCLF) consistently outperformed the single-task baseline (T-CLF) across all evaluation metrics. Notably, the F1-score increased from 0.660 to 0.714, and the AUC rose from 0.830 to 0.876. This performance gain is attributed to the inductive bias from the auxiliary tasks (casualties, property losses, pollution), which functions as an

effective regularization mechanism. By learning a shared representation for all outcomes, the model is constrained to capture the fundamental factors underlying accident events rather than superficial, task-specific patterns, leading to a more robust and generalizable model. Furthermore, the E-MCLF model, which integrates a generative component, advanced performance even further, achieving an AUC of 0.943. This demonstrates that jointly optimizing the generative and discriminative objectives within a multi-task framework can further refine the shared representations, leading to state-of-the-art predictive accuracy.

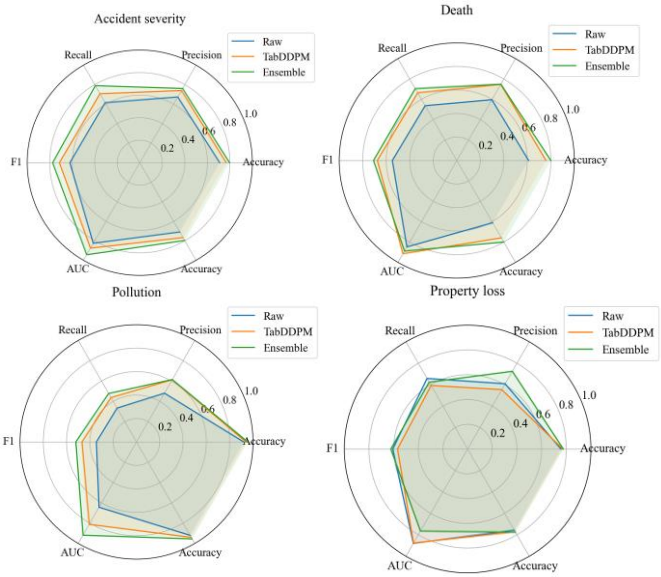


Fig. 14. Comparison of prediction results of each task under different synthesis methods.

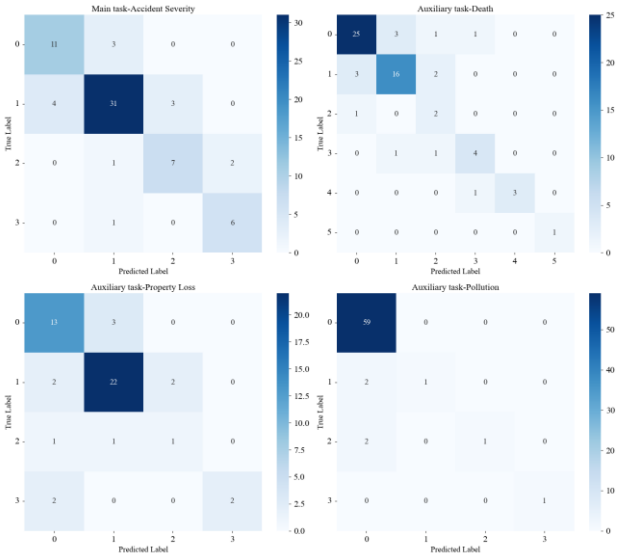


Fig. 15. Confusion matrix of multi-task prediction results of our model.

To elucidate the mechanisms underlying the success of our multi-task framework,

we analysed the model's performance on the auxiliary tasks. As shown in Fig. 14, the model achieved high F1-scores and AUC values across these tasks, indicating that the learned shared representation effectively captures features relevant not only to the primary task but also to related sub-tasks.

A granular analysis of the confusion matrices provides further insight (Fig. 15). For the primary task (accident severity), our model's confusion matrix exhibits a strong diagonal dominance, confirming high classification accuracy. The model also demonstrates high precision and recall for the majority classes in the auxiliary tasks. Notably, the near-perfect classification for the non-pollution class (59/59 correct) suggests that pollution-related features are highly discriminative and were effectively leveraged within the shared representation. These results validate that the auxiliary tasks provide synergistic information that enriches the shared feature space. The model's proficiency across diverse outcomes demonstrates the efficacy of the proposed architecture, where a shared encoder learns a holistic representation of the problem domain. This approach fundamentally contrasts with training isolated single-task models, which fail to exploit inter-task correlations, thereby limiting the model's generalization capability.

4.6 Causal pathways verification (RQ3)

While the preceding sections have established the superior predictive performance of our E-MCLF model, a more profound question remains (RQ3): Does the learned latent representation Z_α function as a genuine causal mediator in the $\mathrm{X} \rightarrow \mathrm{Z}\alpha \rightarrow \mathrm{Y}$ pathway, or is it merely a powerful statistical correlate? This section aims to empirically validate this causal hypothesis by employing a multi-faceted approach grounded in causal inference principles. We validate three hypotheses: H1: Causal mediation hypothesis, H2: Mediation necessity and sensitivity hypothesis, and H3: Disentanglement hypothesis. A positive validation would imply that our model has not only learned what features are predictive but has also uncovered a plausible underlying mechanism of how structured features lead to accident severity through the lens of unsafe human/organizational factors.

4.6.1 Disentanglement Verification

To test our hypotheses, we employ a counterfactual-based causal mediation analysis (CMA) framework to decompose the Total Effect (TE) of inputs X on the prediction Y into a Natural Direct Effect (NDE) and a Natural Indirect Effect (NIE). Specifically, to test H1 (Causal Mediation), we quantify the NIE transmitted through

the causal mediator Z_α, and calculate the Mediation Proportion ($MP = \frac{NIE}{TE}$); a statistically significant NIE and a high MP value provide strong support for this hypothesis. To test H3 (Disentanglement), we conduct a comparative analysis by computing the pseudo-NIE that passes through the non-causal channel Z_β. Successful disentanglement is demonstrated when the effect from the causal channel is dominant and the effect from the non-causal channel is negligible ($NIE(Z_\alpha) \gg NIE(Z_\beta)$).

To ensure statistical robustness across all analyses, we use non-parametric bootstrapping with 2000 resamples to compute 95% confidence intervals (CIs) for all estimated effects. An effect is deemed statistically significant if its 95% CI does not contain zero.

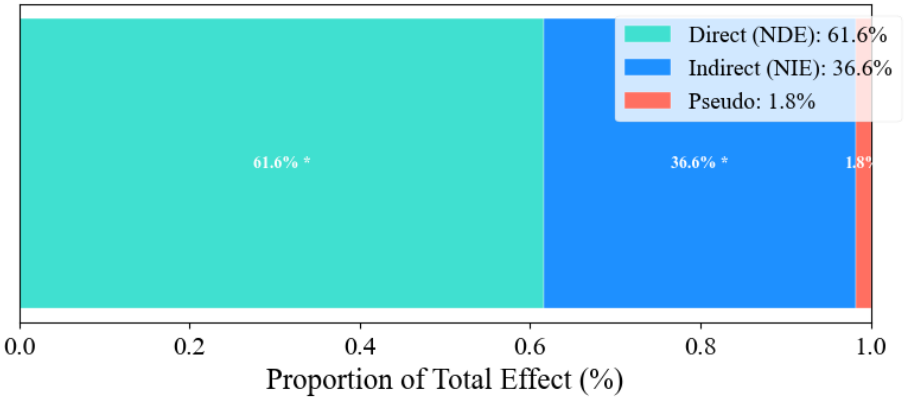


Fig. 16. Decomposition of the total causal effect.

The causal mediation analysis provides compelling evidence for successful disentanglement, robustly validating both our primary hypotheses (H1 and H3). In support of H1, the analysis confirms that the learned factors Z_α act as significant mediators. This indirect effect is statistically significant (95% CI does not contain zero), with a Mediation Proportion (MP) of 0.366, indicating that over one-third of the model's predictive reasoning is channeled through these structured, high-level concepts. Critically, the analysis also validates H3 by demonstrating a successful separation of causal and non-causal pathways. A stark difference is observed between the effect mediated by the causal channel Z_α (NIE = 36.6%) and the negligible pseudo-effect from the non-causal channel Z_β (Pseudo-NIE = 1.8%). This disparity, where the causal pathway's influence is over 20 times greater, confirms that the model has learned a meaningful causal representation while effectively isolating and neutralizing the statistically insignificant influence of non-causal factors.

4.6.2 Mediation Ablation Study

To rigorously validate the functional contribution of the learned mediator Z_α, we conducted a mediation ablation study designed to quantify its impact on the final prediction. We systematically corrupted the information flow through the Z_α channel by randomly permuting an increasing fraction $\rho \in [0,1]$, of its feature dimensions, where $\rho = 0$ represents the original mediator and $\rho = 1$ signifies complete destruction of its informational content.

The resulting impact was measured by the degradation in model performance (Accuracy and AUC drop) and the shift in output distribution (KL divergence). As illustrated in Fig. 17, the results demonstrate the critical role of Z_α. We observe a clear positive correlation between the permutation ratio ρ and performance degradation; as ρ increases, both the Accuracy drop and AUC drop show a distinct upward trend, culminating in substantial drops of approximately 0.075 and 0.08, respectively, when the channel is fully ablated ($\rho = 1$). This confirms that the information encoded in Z_α is necessary for the model to achieve its high predictive accuracy. Concurrently, the KL divergence between the original and perturbed prediction distributions also grows in tandem with ρ, indicating that the model's predictions are highly sensitive to the integrity of the information within the Z_α channel, as even partial perturbations lead to significant shifts in the model's output. Collectively, these ablation results provide compelling evidence that Z_α is not a redundant feature but a functionally indispensable component of the model's decision-making process.

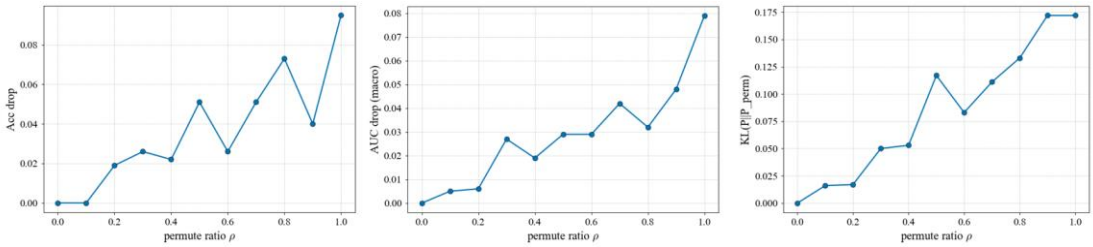


Fig. 17. Functional necessity and sensitivity of the mediator via controlled perturbation.

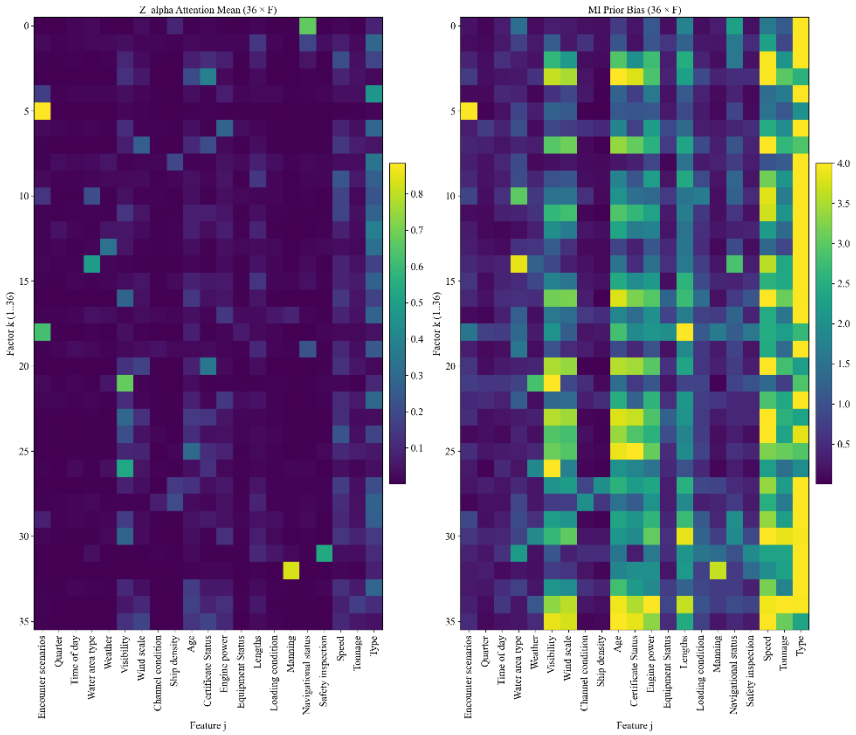


Fig. 18. Alignment between learned Attention and the MI Prior.

To validate the successful learning of an interpretable and disentangled latent representation Z_α, we visualize and compare the model's final attention mechanism against the guiding MI prior. The results, presented in Fig. 18, provide compelling evidence for this claim. The MI Prior Bias (right panel) illustrates the intended structured mapping, where specific latent factors (k, y-axis) are encouraged to capture information from distinct input features (j, x-axis). The learned Z_α Attention Mean (left panel) demonstrates two critical properties: first, it is highly sparse, confirming that each latent factor has learned to specialize by focusing on a minimal subset of input features rather than creating an entangled mixture of signals. Second, the locations of these high-attention weights (bright spots) align with remarkable fidelity to the guiding prior, indicating that the training process successfully steered the model to learn the desired feature-to-factor associations. For instance, the factors assigned to capture concepts like "Encounter scenarios" (k≈5), "Wind scale" (k≈20), and "Navigational status" (k≈32) have precisely isolated these signals as intended. This verified disentanglement is fundamentally important, as it ensures that each dimension of Z_α represents a clean, isolated concept, thereby providing a robust and trustworthy foundation for the subsequent Causal Mediation Analysis.

Table 5 Top-K metrics and MRR results.

Metric	value
Hit@1	0.5668

Hit@3	0.7762
Hit@5	0.8303
Hit@10	0.9634
MRR	0.7842

To evaluate the model's ability to accurately identify and rank the true unsafe factors from a set of candidates, we present top-K ranking metrics and the Mean Reciprocal Rank (MRR) in Table 5. The model achieves a Hit@1 of 0.5668, correctly identifying the ground-truth unsafe factor as its top prediction in 56.7% of cases. Performance improves substantially when considering more candidates, with Hit@10 reaching 0.9634, which signifies that the true factor is captured within the top 10 predictions over 96% of the time. The high MRR of 0.7842 further confirms that, on average, the correct factors are ranked very near the top.

Collectively, these results highlight a key strength of our model: while it may not always achieve perfect top-1 accuracy, it functions as an exceptionally effective candidate-filtering and ranking engine for risk factors. It demonstrates a powerful ability to narrow down a wide range of possibilities to a very small, highly relevant set of top-ranked predictions. This capability is particularly valuable in practical applications where the model can serve as a highly reliable first-stage filter, allowing human experts or subsequent systems to focus their attention on a manageable number of high-risk scenarios.

4.7 Total Contribution and Individual Effect Analysis (RQ4)

4.7.1 Overall effect analysis

To understand the average causal contribution of each unsafe factor at a macro level, we first used a factor masking method to calculate the average probability change of each factor for the four accident consequences at different levels. In Fig. 19, a positive value indicates that the presence of the factor increases the probability of that consequence level (risk promotion), while a negative value indicates a decrease in the probability (risk suppression).

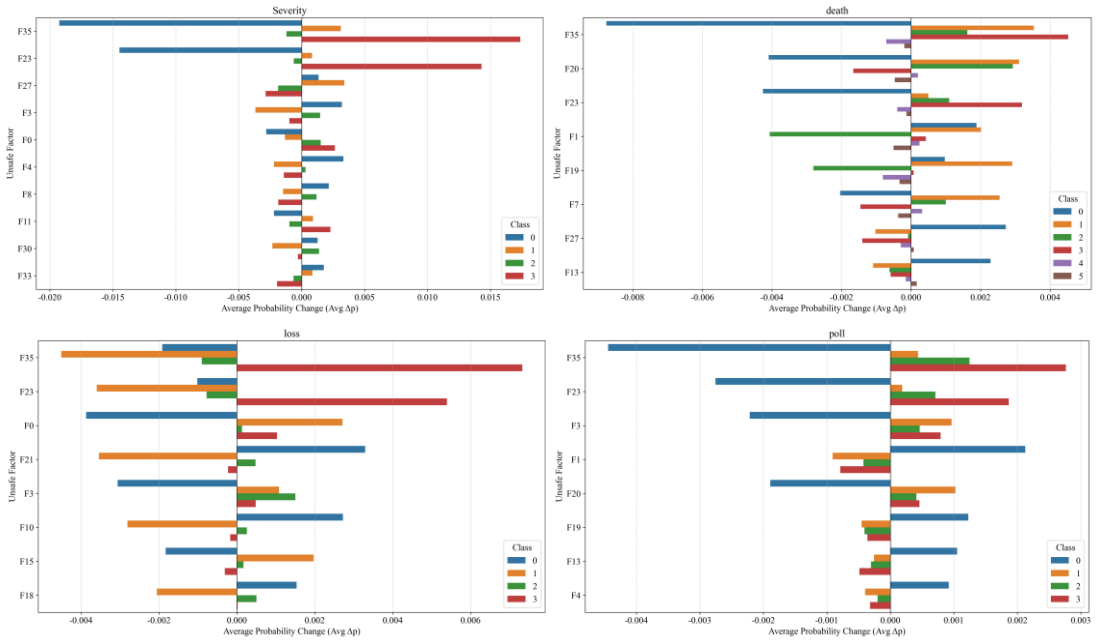


Fig. 19. Class-wise contribution of unsafe factors to the Model's Predictions across four tasks.

As shown in Fig. 19, factors F35 and F23 are the most critical sources of systemic risk. Across all four outcomes analysed (severity, losses, death, pollution), they show consistent and strong negative impacts, significantly increasing the probability of the most severe outcomes. The model successfully captures the unique associations between different unsafe factors and specific outcomes. This means that not all errors lead to the same disaster. Certain operational errors are more likely to lead to specific types of severe consequences, providing a basis for targeted prevention. In the two tasks of severity and property loss, the impact of F35 and F23 is the most prominent. They are the core reasons that directly push accidents from "minor" to "critical" level, or from "low loss" to "huge loss". In the death task, in addition to the F35 and F23, the F20 is also a key factor leading to high death levels. In addition, factors such as F1 and F19 are more strongly associated with causing a small number of deaths (1-2 people), revealing the differences in lethality between different factors. Among the pollution task, F35 and F23 remain the primary causes of severe contamination. Interestingly, while F1 (improper use of navigation equipment) is a contributing factor, its results tend to reduce the probability of severe contamination and increase the likelihood of minor contamination. This suggests that the types of accidents caused by this factor are less likely to result in large-scale leaks, perfectly demonstrating causal specificity.

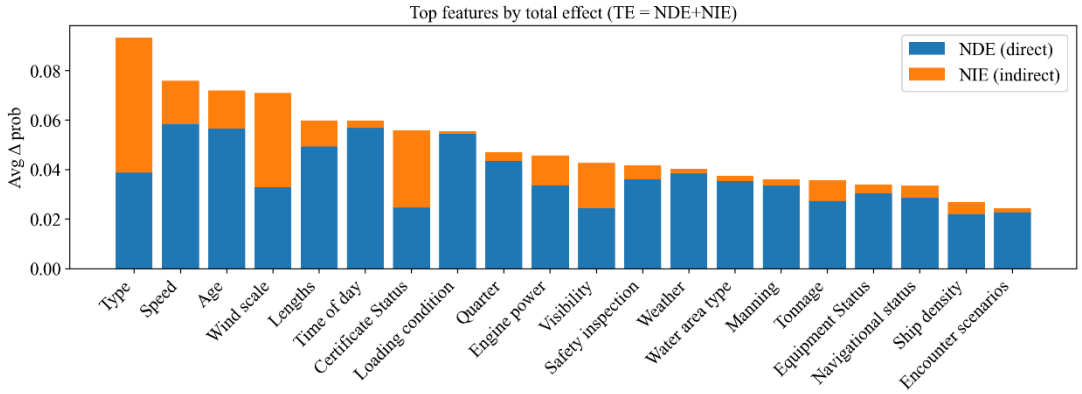


Fig. 20. Decomposition of total causal effects of features into NDE and NIE Effects.

Fig. 20 shows the total causal effect of the original input features, decomposed into NDE and NIE. For the vast majority of top-level influencing features, such as vessel type, speed, wind scale, and age, their indirect effects (NIE, shown in orange) account for a significant portion of the total effect. This strongly suggests that these external conditions did not directly cause the accident, but rather triggered certain key "potentially unsafe conditions" (for example, high speed combined with inclement weather leading to a "loss of control"), which in turn led to serious consequences. This provides solid empirical support for our proposed causal mediation hypothesis. Different features have different effect paths. For example, environmental factors such as weather, visibility, and wind scale largely exert their influence through indirect effects (NIE). This is highly consistent with intuition: inclement weather itself does not directly cause losses, but rather exerts its influence through mediating factors such as vessel maneuverability and crew perception. Management factors such as Certificate Status, on the other hand, have high NDEs, indicating that the violation itself constitutes a direct source of risk.

4.7.2 Individual Effect Analysis

To validate the effectiveness, accuracy, and interpretability of our proposed multi-task causal learning framework (MCLF) in complex real-world scenarios, we selected a publicly available, untrained Spanish maritime investigation report (CIAIM-28/2015) as a test set. Tools such as causal diagrams, factor masking effects, and natural direct/indirect effect (NDE/NIE) decomposition are used to achieve transparent and logical attribution analysis.

On July 20, 2015, near Tarifa, Spain, the high-speed vessel "Maria Dolores" collided with the fishing vessel "Rinconcillo" in dense fog. The fishing vessel passed through the passage between the ro-ro passenger ferry's catamaran hulls, causing

damage to the vessel's superstructure and equipment. The accident resulted in no casualties, no marine pollution, and minimal property damage.

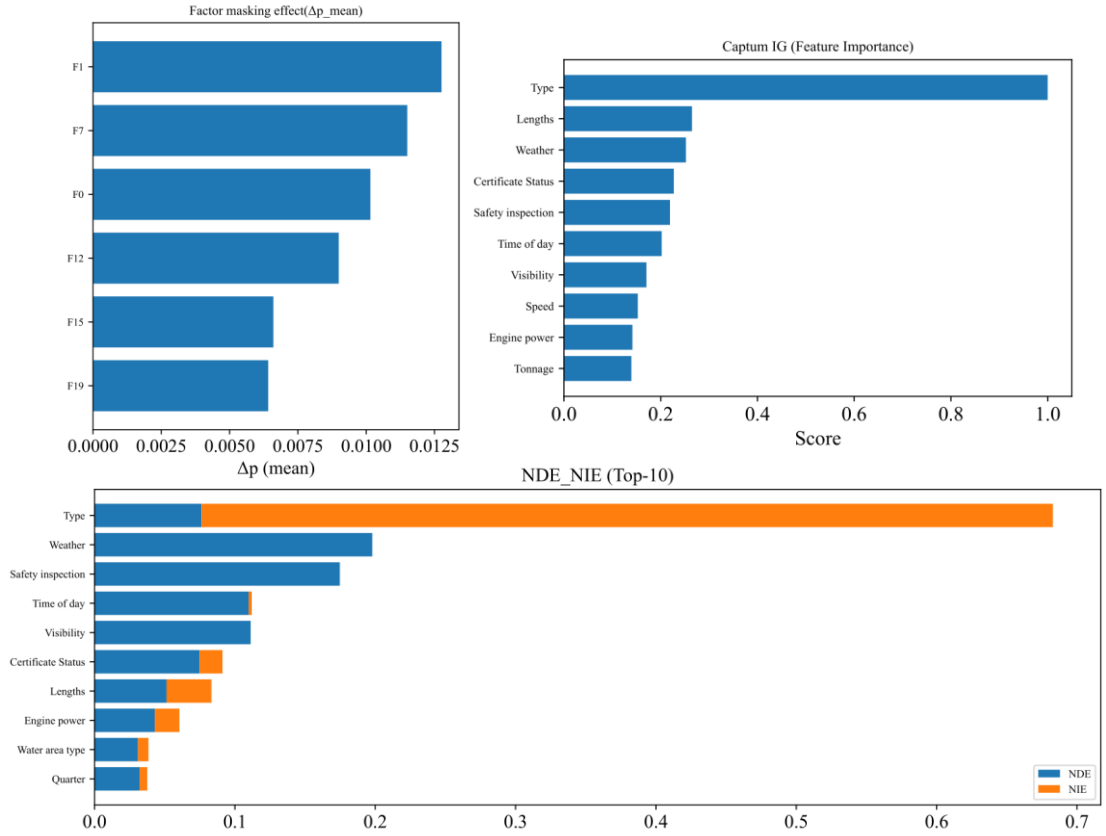
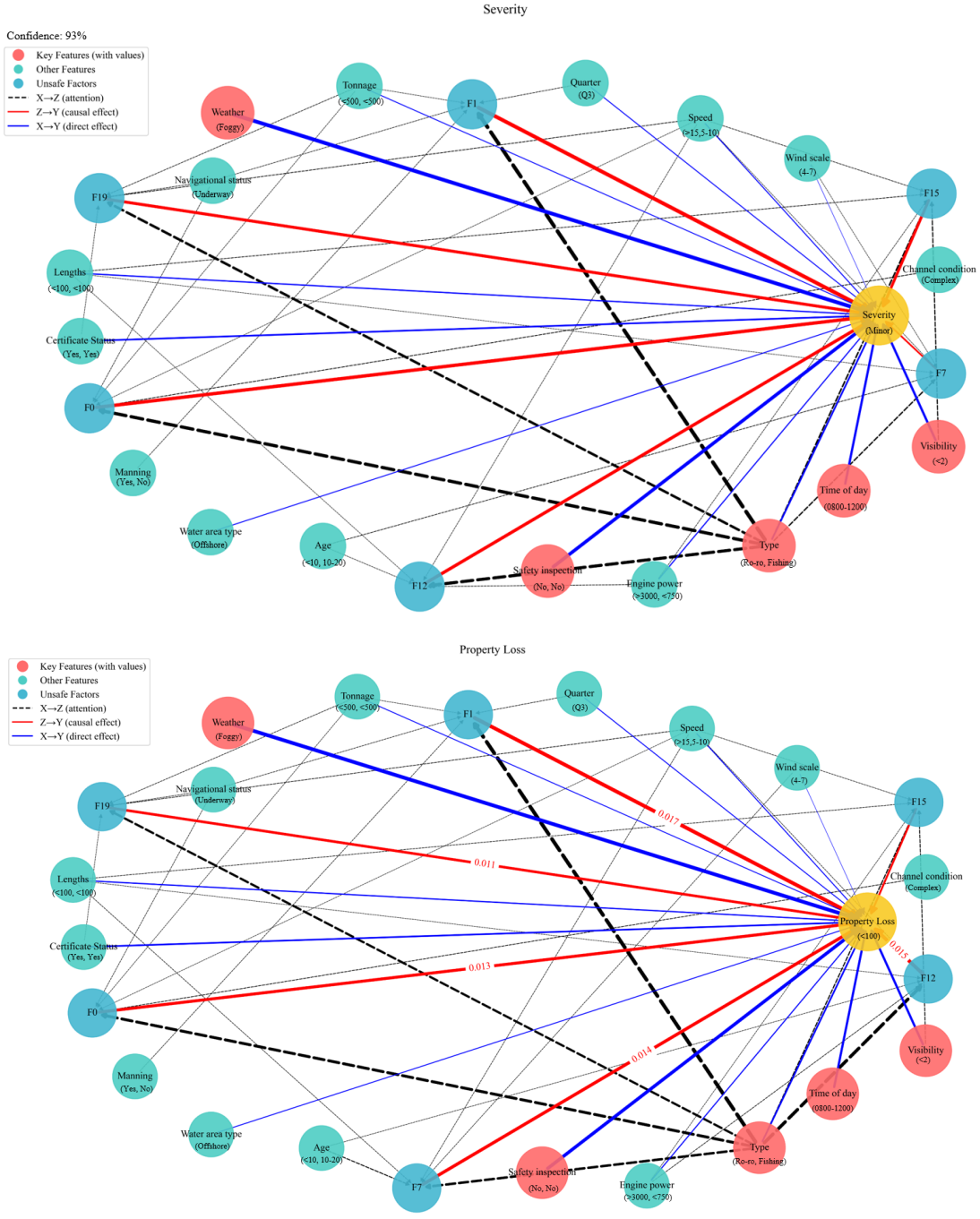


Fig. 21. Comprehensive evaluation of feature importance and causal effects.

As shown in Fig. 21, the model, through a cross-attention mechanism, identified the key human errors (mediating factors) that contributed to the accident in this specific scenario. Based on the Δp mean plot, the model identified F1 (improper use of navigation equipment) as the most critical factor, which aligns perfectly with the report's core finding: "The captain's improper switching of radar ranges led to target loss." The model identified F7 (misjudgment of the situation) and F0 (negligence of lookout) as the second and third most important factors. The report confirmed that, despite the risk of collision, "no one on board perceived the risk" (F7), and the chief officer failed to report the radar target (F0). The remaining factors (F12, F15, and F19) are all clearly supported by evidence in the report, including the fishing vessel's incorrect collision avoidance maneuver (F19), a lack of communication on the bridge (F15), and overall irregularities in operation (F12). The model's identification of key mediating factors is consistent with the in-depth investigation findings of maritime experts, strongly demonstrating the model's inference capabilities. The Vessel Type feature exhibits minimal NDE and significant NIE, indicating that the "Ro-Ro ship vs.

fishing vessel" ship type combination itself did not directly cause the accident. Instead, this complex encounter scenario with a high speed and tonnage difference greatly increased the likelihood of crew errors (such as F0, F1, and F7), which indirectly contributed to the accident. Both the Weather and Visibility features exhibit significant NDE and NIE. Bad weather and low visibility are both hazards in themselves and catalysts, increasing the likelihood of crew operational errors and misjudgments, which in turn exacerbate the accident indirectly. The DE/NIE decomposition diagram clearly quantifies how different features influence the accident, intelligently identifying which features are direct hazards and which contribute indirectly by amplifying unsafe factors.



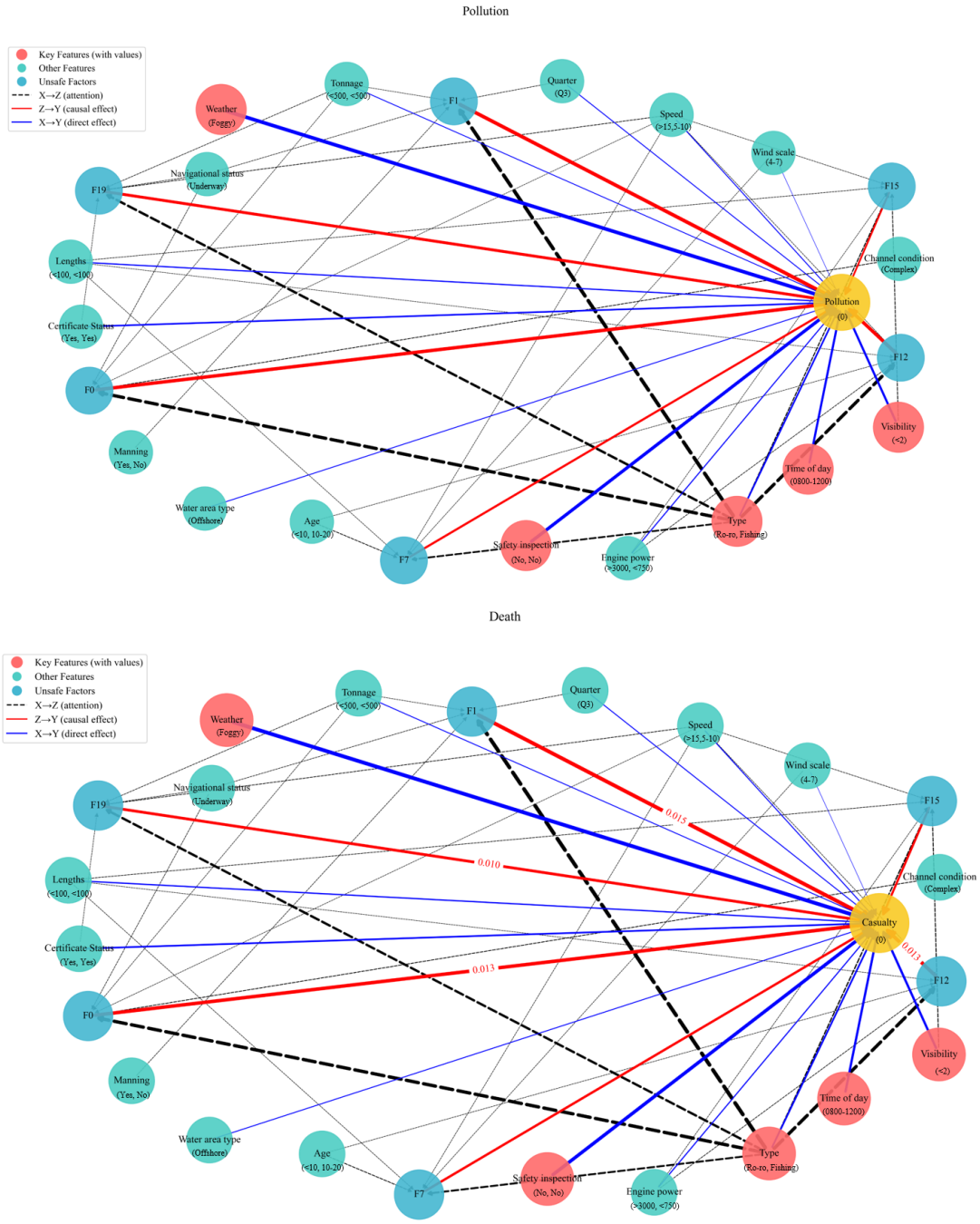


Fig. 22. The causal pathways diagrams for different tasks.

As shown in Fig. 22, the four graphs (Death, Severity, Property Loss, and Pollution) correspond to the four subtasks in multi-task learning. They share much of the same structure but may differ in the weighting of details, reflecting the model's differing attributions of different outcomes. The accident severity was rated as minor, with a 93% confidence level. The prediction of the accident's consequences was accurate. The model identifies that the "intersecting encounter between the ro-ro ship (high speed, large tonnage) and the fishing vessel (slow speed)" is an extremely complex and high-risk scenario. The thickest black dotted line indicates that the model believes that this

scenario greatly increases the possibility of crew members making mistakes in F1, F19, and F7. This perfectly confirms the conclusion in the NDE/NIE analysis that "Type has a huge indirect effect." Fog and low visibility, two strongly correlated environmental factors, will directly lead to a sharp increase in the risk of F0, while forcing crew members to rely more on and possibly misuse navigation equipment, thereby activating the risk of F1. F7's contribution to property damage is greater than its contribution to human fatalities. Behaviors such as F0 and F19 that directly lead to physical collisions are more closely related to casualties.

Through the visual analysis of characteristics and factor effects of individual cases, not only the explanatory power of the model is verified, but also the heterogeneity of risk factors, nonlinear action mechanisms and intrinsic correlations between consequences in the evolution of accidents are revealed.

4.8 Robustness analysis

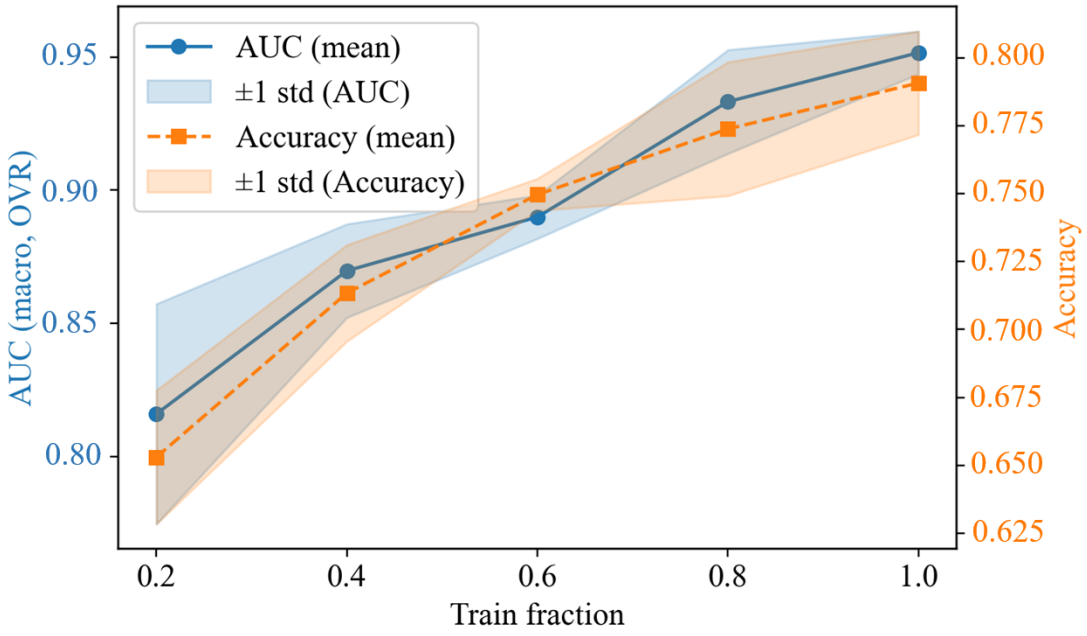


Fig. 23. Learning curves of model performance vs. training data size

The model's data efficiency and generalization ability were evaluated using learning curves (see Fig. 23). The Fig. 23 illustrates how model performance (macro-average AUC and accuracy) improves as the proportion of the training set, which includes synthetically generated samples, increases. Both metrics exhibit a steep, monotonic rise, demonstrating remarkable learning efficiency. A notable finding is that even with only 40% of the training data, the model achieves a high macro-average AUC of approximately 0.87. This high sample efficiency is a direct result of the synergy between our data synthesis strategy and model architecture. The synthetic data enriches

the feature space and mitigates class imbalance, allowing the model to capture key causal relationships from a limited set of real samples. This is further enhanced by the MI prior-guided attention and causal decoupling mechanism, which constrains the learning space, guiding the model to focus on core causal pathways and avoid overfitting, thus enabling rapid convergence to a high-quality solution.

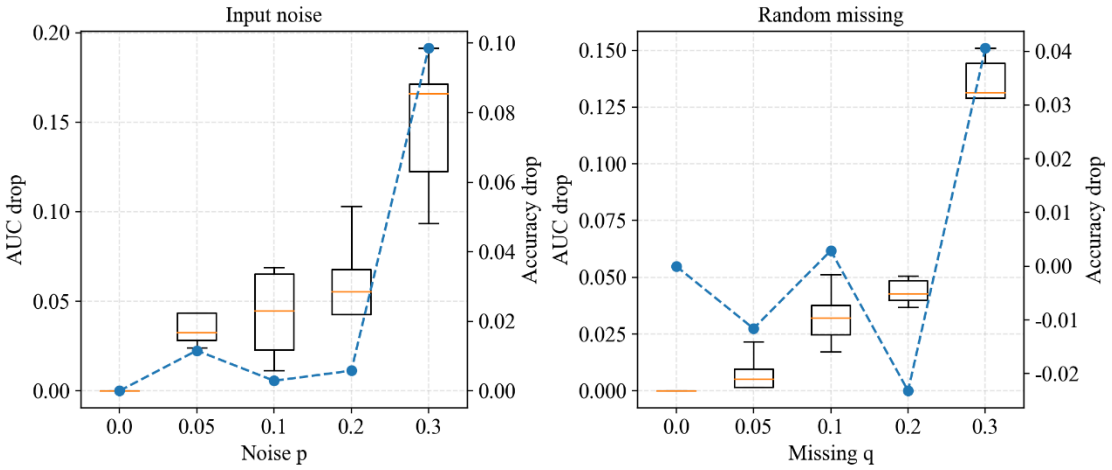


Fig. 24. The performance degradation of the model in the presence of input feature noise and random missing features

The robustness of the model was tested by introducing controlled perturbations. Fig. 24 shows the performance degradation of the model in the presence of input feature noise (left) and random missing features (right). In the noise experiment, we randomly replaced the category values of some features with probability p to simulate errors in data collection or recording. The results show that the model exhibits graceful performance degradation. When the noise level increases from 0 to 20%, the performance degradation is extremely limited, with the median AUC drop remaining within 0.05, demonstrating its strong resilience to feature perturbations. The model demonstrates even greater robustness in the face of missing data. When the feature missing ratio q increases from 0 to 30%, the model's performance remains remarkably stable. In particular, under the challenging condition of 20% missing features, the median AUC drop remains below 0.05, with little loss in accuracy. This demonstrates that our model architecture can effectively handle incomplete information, which is crucial for real-world applications such as maritime safety. The mechanism behind this is that the attention-based intermediate inference process allows the model to dynamically reallocate the importance weight of a specific feature to other relevant, observable features when the feature is missing, thereby effectively compensating for the missing information in the context.

4.9 Ablation Study

To rigorously validate the necessity and specific contribution of each key component within our proposed framework, we conducted a comprehensive ablation study. The objective of this study extends beyond merely quantifying performance degradation; it aims to empirically demonstrate how each module contributes to the model's core function. We designed several variants of our full model by systematically removing one component at a time. The Corr indicator is introduced to represent the correlation between the direct path and the indirect path representation, and the lower the better.

Table 6 Ablation experiment results.

Model Variant	Accuracy	Precision	Recall	F1	AUC	Corr (↓)
E-MCLF(our model)	0.797	0.761	0.790	0.774	0.943	0.088
w/o GRL	0.801	0.763	0.795	0.786	0.946	0.101
w/o Ortho. Loss	0.789	0.752	0.781	0.765	0.938	0.324
w/o Perm. Consist.	0.782	0.745	0.775	0.758	0.933	0.105

The most striking result emerges from the removal of the Orthogonal Loss (w/o Ortho. Loss). In this variant, the value surged dramatically from 0.088 to 0.324, an almost threefold increase. This indicates a severe entanglement between the causal and confounding representations, rendering the decoupling mechanism ineffective. This fundamental failure in representation separation directly undermined the model's predictive power, leading to significant drops in both AUC (to 0.938) and F1-score. By imposing a strong geometric constraint, it fundamentally enforces the independence of the two representations. The removal of the Gradient Reversal Layer (w/o GRL) reveals a more nuanced trade-off. Although its ablation led to a marginal improvement in predictive metrics like AUC (0.946) and F1-score (0.786), the representation correlation concurrently deteriorated to 0.101. This phenomenon suggests that GRL, through adversarial training, successfully purifies the causal representations by preventing the model from exploiting spurious cues present in the confounding features. While leveraging these spurious correlations might offer a slight performance gain on this specific test set, it compromises the model's generalization and interpretability. Finally, ablating the Permutation Consistency module (w/o Perm. Consist.) resulted in a comprehensive decline in performance. This clearly demonstrates that Permutation Consistency, acting as an effective regularization technique, enhances the model's

generalization and prediction accuracy by forcing it to learn stable relationships insensitive to feature perturbations.

In conclusion, these three components each play a distinct yet complementary role. The complete E-MCLF model demonstrates a superior balance across all metrics, with an AUC of 0.943 and a low Corr of 0.088, attesting to the advanced design and synergistic effect of its architecture.

5. Conclusion

This study proposes a novel causal representation framework that transcends traditional correlation analysis, revealing the dynamic and heterogeneous causal mechanisms underlying accident outcomes. Our primary scientific contribution lies in the empirical discovery and quantification of dynamic causal mediation effects. We demonstrate that pre-accident conditions (such as weather and vessel type) exert their influence on accident severity not through a singular direct effect, but via a complex pathway where direct effects coexist with indirect mediated effects. Their impact is transmitted not only directly to accident consequences but also indirectly through a diverse array of latent unsafe factors. Crucially, the significance of these mediating pathways dynamically varies according to the specific context of each accident. Building on this, we demonstrate the profound analytical power of decomposing causal pathways across multi-dimensional consequences. Our parallel multi-task framework successfully isolated distinct causal signatures for different types of harm from the same event data. Our analysis found that pathways leading to death are driven by a combination of deep-seated systemic failures (such as inadequate safety management and equipment failure) and critical operational errors (such as negligent lookout). In stark contrast, property loss is more directly attributable to equipment failure itself and misjudgment of the situation. Meanwhile, severe pollution events exhibit their own unique causal pattern, typically linked to specific high-risk vessel types (like oil tankers, bulk carriers, and even fishing vessels) and the combined effect of multiple systemic risks. Collectively, these findings validate a paradigm shift from homogeneous, post-hoc investigation to heterogeneous, pre-accident attribution. The ability to generate granular, outcome-specific, and individualized causal narratives provides a level of mechanistic insight previously obscured by traditional, single-consequence models.

While the model demonstrates significant advantages, it also has limitations. As a data-driven model, the validity of its conclusions is highly dependent on the quality and completeness of accident data. Recording biases or missing information can affect the

accuracy of causal relationships. Furthermore, the model struggles to fully control for potential unobserved confounding factors (such as crew psychological states and organizational management culture). Future research directions include, first, further incorporating real-time data streams (such as AIS ship dynamics data and meteorological monitoring data) to enhance the model's predictive and dynamic updating capabilities, enabling real-time identification of high-risk areas and time periods. Second, natural language processing methods can be combined to mine implicit information within accident report texts, enriching the input dimensions of the causal network. In summary, this study represents a key step forward in expanding the multidimensional quantitative analysis of accident consequences. While limitations remain, the proposed causal representation framework provides a solid foundation for future in-depth research and practical applications.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grant Nos. 52171353 and 52471387), the Fundamental Research Funds for the Central Universities (No. 24CX02024A) and the Open Project of the State Key Laboratory of Chemical Safety, the open project of State Key Laboratory of Maritime Technology and Safety, Wuhan University of Technology (No. 29-19-2). This research has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement H2020- MSCA-IF2018–840425.

Appendix

Table 1 Candidate contributing factors to ship collision.

Factors	RIFs	Unit	States
Ship factors:	Lengths	m	<100、100-150、151-200、>200
	Tonnage	tons	<500、500-3000、>3000
	Engine power	kW	<750, 750–3000, >3000
	Ship age	years	<10, 10–20, >20
	Speed	knots	<5, 5–10, 11–1, >1
	Ship type	/	General cargo, Multipurpose, Bulk carrier, Engineering vessel, Sand carrier, Dry cargo, Passenger, Ro-ro, Tugboat, Fishing, Container ship, Oil tanker, Others
	Loading condition	/	Normal load, Ballast, Overloaded
	Navigational status	/	Underway, Anchored, Moored, Operating
	Equipment status	/	Normal, Abnormal
	Manning	/	Compliant, Non-compliant (minimum safe manning)
	Certificate status	/	Yes, No
	Safety inspection	/	Yes, No
Environmental factors:	Quarter	/	Q1, Q2, Q3, Q4
	Time of day	/	00:00–04:00, 04:00–08:00, 08:00–12:00, 12:00–16:00, 16:00–20:00, 20:00–24:00
	Water area type	/	Offshore, Channel, Port, Anchorage, River, Estuary, Fishing zone, Bridge area
	Weather	/	Sunny, Cloudy, Rainy, Foggy
	Visibility	km	<2, 2–5, 6–1, >10
	Wind scale	level	<4, 4–7, >7
	Channel condition	/	Good, Moderate, Complex
	Ship density	/	High, Low
Encounter Scenarios:	Encounter scenarios	/	Crossing, Head-on, Overtaking

Table 2 Collision accident consequence variables.

Dependent Variable	Number of Categories	Description
Death severity	6	0:n=0
		1:n=1
		2:n=2
		3:n=3-5

			4:n=6-9 5: $n \geq 10$
Property loss severity	4		0: $n < 100$ 1: $100 \leq n < 1000$ 2: $1000 \leq n < 5000$ 3: $5000 \leq n$
Severity level	4		0:n=Minor 1:n=General 2:n=Primary 3:n=Critical
Pollution	4		0:n=0 1: $0 < n < 100$ 2: $100 \leq n < 500$ 3: $500 \leq n$

Table 3 Frequency statistics of the 36 risk factors.

Variables	Risk factors	Frequency (%)	Variables	Risk factors	Frequency (%)
F0	Lookout negligence	79.65	F18	Improper timing selection	0.14
F1	Ineffective use of navigation aids	28.20	F19	Improper collision avoidance action	42.14
F2	Unused safe speed	30.35	F20	No avoidance measures were taken	59.74
F3	Failure to detect the ship early	47.09	F21	No foggy navigation measures were taken	2.47
F4	Illegal crossing	4.50	F22	Unchecked the effectiveness of collision avoidance actions	4.65
F5	Illegal overtaking	6.10	F23	Equipment Failure	25.30
F6	Risky navigation	2.18	F24	Not equipped with navigation aids	14.24
F7	Situational Misjudgment	51.29	F25	Unseaworthiness	21.22
F8	Improper duty arrangements	29.88	F26	Poor visibility	25.60
F9	Improper ship operation	17.30	F27	High traffic density	19.91
F10	Obstructing navigation of	6.40	F28	Complex navigational environment	27.61

	other vessels				
F11	Weak safety awareness	16.72	F29	Influence of wind, waves and currents	15.00
F12	Failure to exercise good Seamanship	11.92	F30	Adverse weather	5.23
F13	Improper display of signals	16.42	F31	Navigation beyond approved area	3.49
F14	Improper anchoring	4.21	F32	Uncertificated crew	29.07
F15	Poor communicatio n	32.12	F33	Insufficient manning	18.16
F16	Lack of navigational skills	11.48	F34	Insufficient knowledge/experience/trainin g	21.66
F17	Failure to navigate with caution	16.28	F35	Inadequate safety management	34.90

References

[1] Cao W, Wang X, Feng Y, Zhou J, Yang Z. Improving maritime accident severity prediction accuracy: A holistic machine learning framework with data balancing and explainability techniques. Reliab Eng Syst Saf. 2025:111648.

[2] Qiao W, Huang E, Zhang M, Ma X, Liu D. Risk influencing factors on the consequence of waterborne transportation accidents in China (2013–2023) based on data-driven machine learning. Reliab Eng Syst Saf. 2025;257:110829.

[3] Jia Q, Fu G, Xie X, Xue Y, Hu S. Enhancing accident cause analysis through text classification and accident causation theory: A case study of coal mine gas explosion accidents. Process Saf Environ Prot. 2024;185:989-1002.

[4] Chen D, Pei Y, Xia Q. Research on human factors cause chain of ship accidents based on multidimensional association rules. Ocean Eng. 2020;218:107717.

[5] Meng X, Li H, Zhang W, Zhou X-Y, Yang X. Analyzing risk influencing factors of ship collision accidents: A data-driven Bayesian network model integrating physical knowledge. Ocean Coastal Manage. 2024;256:107311.

[6] Puisa R, Lin L, Bolbot V, Vassalos D. Unravelling causal factors of maritime incidents and accidents. Saf Sci. 2018;110:124-41.

[7] Zhang X, Chen P, Mou J, Chen L, Li M. Critical causation factor analysis in ship collision accidents with complex network. Ocean Eng. 2025;315:119837.

[8] Feng Y, Wang X, Chen Q, Yang Z, Wang J, Li H, et al. Prediction of the severity of marine accidents using improved machine learning. Transp Res Part E Logist Transp Rev. 2024;188:103647.

[9] Lan H, Ma X, Qiao W, Deng W. Determining the critical risk factors for predicting the severity of ship collision accidents using a data-driven approach. Reliab Eng Syst Saf. 2023;230:108934.

[10] Gan L, Gao Z, Zhang X, Xu Y, Liu RW, Xie C, et al. Graph neural networks enabled accident causation prediction for maritime vessel traffic. Reliab Eng Syst Saf. 2025;257:110804.

[11] Li S, Pu Z, Cui Z, Lee S, Guo X, Ngoduy D. Inferring heterogeneous treatment effects of crashes on highway traffic: A doubly robust causal machine learning approach. Transp Res Part C Emerging Technol. 2024;160:104537.

[12] Fu S, Wu M, Zhang Y, Zhang M, Han B, Wu Z. Coupling and causation analysis of risk influencing factors for navigational accidents in ice-covered waters. Ocean Eng. 2025;320:120280.

[13] Meng X, Li H, Zhang W, Zhou X-Y, Yang X. Analyzing ship collision accidents in China: A framework based on the NK model and Bayesian networks. Ocean Eng. 2024;309:118619.

[14] Fu S, Yu Y, Chen J, Xi Y, Zhang M. A framework for quantitative analysis of the causation of grounding accidents in arctic shipping. Reliab Eng Syst Saf. 2022;226:108706.

[15] Ma L, Ma X, Lan H, Liu Y, Deng W. A data-driven method for modeling human factors in maritime accidents by integrating DEMATEL and FCM based on HFACS: A case of ship collisions. Ocean Eng. 2022;266:112699.

[16] Chen J, Zhuang C, Shi J, Jiang H, Xu J, Liu J. Risk factors extraction and analysis of Chinese ship collision accidents based on knowledge graph. Ocean Eng. 2025;322:120536.

[17] Guo Y, Ai X, Luo W. A multi-task learning risk assessment method for the chemical process industry. Process Saf Environ Prot. 2024;186:980-94.

[18] Yang Z, Zhang W, Feng J. Predicting multiple types of traffic accident severity with explanations: A multi-task deep learning framework. Saf Sci. 2022;146:105522.

[19] Chen P, Zhang A, Zhang S, Dong T, Zeng X, Chen S, et al. Maritime Near-Miss prediction framework and model interpretation analysis method based on Transformer neural network model with multi-task classification variables. Reliab Eng Syst Saf. 2025;257:110845.

[20] Yoon J. Prediction of high-risk areas using the interpretable machine learning: Based on each determinant for the severity of pedestrian crashes. J Transp Geogr. 2025;126:104216.

[21] Parsa AB, Movahedi A, Taghipour H, Derrible S, Mohammadian AK. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accid Anal Prev. 2020;136:105405.

[22] Arif S, MacNeil MA. Applying the structural causal model framework for observational causal inference in ecology. Ecol Monogr. 2023;93:e1554.

[23] Gomes-Franco K, Rivera-Izquierdo M, Martín-delosReyes LM, Jiménez-Mejías E, Martínez-Ruiz V. Explaining the association between driver's age and the risk of causing a road crash through mediation analysis. Int J Environ Res Public Health. 2020;17:9041.

[24] Liu F, Liu W, Liu J, Zhong B, Sun J. Mitigating potential risk via counterfactual explanation generation in blast-based tunnel construction. Adv Eng Inf. 2025;65:103227.

[25] Sun H, Poskitt CM, Sun Y, Sun J, Chen Y. ACAV: a framework for automatic causality analysis in autonomous vehicle accident recordings. Proceedings of the IEEE/ACM 46th International Conference on Software Engineering2024. p. 1-13.

[26] Shelke MS, Deshmukh PR, Shandilya VK. A review on imbalanced data handling using undersampling and oversampling technique. Int J Recent Trends Eng Res. 2017;3:444-9.

[27] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-57.

[28] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. Pacific-Asia

conference on knowledge discovery and data mining: Springer; 2009. p. 475-82.
[29] Maciejewski T, Stefanowski J. Local neighbourhood extension of SMOTE for mining imbalanced data. 2011 IEEE symposium on computational intelligence and data mining (CIDM): IEEE; 2011. p. 104-11.
[30] Sáez JA, Luengo J, Stefanowski J, Herrera F. SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inf Sci. 2015;291:184-203.
[31] Yi X, Xu Y, Hu Q, Krishnamoorthy S, Li W, Tang Z. ASN-SMOTE: a synthetic minority oversampling method with adaptive qualified synthesizer selection. Complex Intell Syst. 2022;8:2247-72.
[32] Li D, Wong YD, Chen T, Wang N, Yuen KF. An ensemble method for investigating maritime casualties resulting in pollution occurrence: Data augmentation and feature analysis. Reliab Eng Syst Saf. 2024;251:110391.
[33] Kotelnikov A, Baranchuk D, Rubachev I, Babenko A. Tabddpm: Modeling tabular data with diffusion models. International conference on machine learning: PMLR; 2023. p. 17564-79.
[34] Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. 2019. arXiv preprint arXiv:190700503. 2019.
[35] Pearl J. Causal inference. Causality: objectives and assessment. 2010:39-58.
[36] Pearl J. Causality: Cambridge university press; 2009.
[37] Tishby N, Pereira FC, Bialek W. The information bottleneck method. arXiv preprint physics/0004057. 2000.
[38] Han H, Wang W-Y, Mao B-H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. International conference on intelligent computing: Springer; 2005. p. 878-87.
[39] Charte F, Rivera AJ, Del Jesus MJ, Herrera F. MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. Knowledge-Based Syst. 2015;89:385-97.
[40] Wang H, Liu Z, Wang X, Graham T, Wang J. An analysis of factors affecting the severity of marine accidents. Reliab Eng Syst Saf. 2021;210:107513.
[41] Liu JZ, Zhu H, Tian F, Chai T, Xue H. Cause analysis of ship collision accident based on complex network theory. International Conference on Internet of Things and Machine Learning (IoTML 2022): SPIE; 2023. p. 235-45.
[42] Chen J, Pu Z, Zheng N, Wen X, Ding H, Guo X. A novel generative adversarial network for improving crash severity modeling with imbalanced data. Transp Res Part C Emerging Technol. 2024;164:104642.