



Data-augmented maritime accident severity prediction: A port state control records-based fusion approach

Tianyi Li ^{a,b}, Xinjian Wang ^{a,b,*}, Chia-Hsun Chang ^b, Manel Grifoll ^c, Hongxiang Feng ^d, Zaili Yang ^{b,e}

^a Navigation College, Dalian Maritime University, Dalian, 116026, PR China

^b Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool, L3 3AF, UK

^c Barcelona School of Nautical Studies, Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona, 08003, Spain

^d Faculty of Maritime and Transportation, Ningbo University, Ningbo, 315211, China

^e School of Civil and Environmental Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore

ARTICLE INFO

Keywords:

Maritime safety
Marine accidents
Port state control
Severity prediction
Machine learning
Conditional tabular denoising diffusion
probabilistic model

ABSTRACT

Data-driven maritime accident analysis based on mainstream accident databases is approaching a high performance ceiling, calling for further breakthroughs from both data quality and modelling perspectives. Port State Control (PSC) inspection records provide additional pre-accident vessel safety information, offering a valuable opportunity to improve accident severity prediction. This study therefore proposes a PSC-assisted accident severity prediction framework that provides a basis for pre-accident assessment. First, a multi-source prediction database is constructed by integrating Paris MoU PSC inspection records with accident information from the Global Integrated Shipping Information System. To address the limited matched samples and high-dimensional categorical features, a two-stage augmentation strategy is developed. Finally, tree-model-based feature selection is applied, and predictive performance is evaluated. The results show that incorporating PSC data reduces the average Mean Squared Error (MSE) by 1.15%. Cross-validation further shows that the complete framework achieves an additional 1.99% reduction in average MSE compared with the PSC-accident fusion-only setting. In particular, the best-performing LightGBM model achieves an MSE reduction of 11.22% relative to the original accident-data-only baseline. This study demonstrates that PSC inspection records are a valuable supplementary data source for maritime accident severity prediction and highlights their significant role in enhancing vessel accident prediction models.

1. Introduction

Maritime safety serves as a critical foundation for the sustainable development of the global shipping industry [1,2]. According to the European Maritime Safety Agency (EMSA), between 2014 and 2023, there were a total of 26,595 maritime accidents [3]. These accidents resulted in 650 fatalities, 7604 injuries, and 602 pollution accidents, including 171 emergency responses to oil spills. These statistics highlight the severe consequences of maritime accidents in terms of human casualties, environmental damage, and economic losses. Although data-driven maritime accident analysis has made substantial progress in recent years, accident prediction based primarily on conventional accident-centred databases is approaching a high performance ceiling, calling for further breakthroughs from both data quality and modelling

perspectives [4–6].

The continued occurrence of serious maritime accidents further suggests that existing accident-centred data sources may still be insufficient for proactive risk identification [7]. This motivates the expansion of maritime accident analysis beyond retrospective accident records toward pre-accident vessel safety information [8,9]. Port State Control (PSC) inspections are an important regulatory mechanism for verifying compliance with international maritime safety and pollution-prevention conventions [10]. PSC records integrate vessel characteristics and regulatory indicators with structured deficiency information covering major safety domains, thereby providing pre-accident vessel safety information that is not fully captured in conventional accident reports [11].

Evidence also supports an association between PSC inspection

* Corresponding author at: Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool, L3 3AF, UK.
E-mail address: X.Wang1@ljmu.ac.uk (X. Wang).

findings and subsequent vessel safety outcomes. Vessels with more inspection deficiencies were reported to have a higher risk of engine-room fires [12], while the “Solong” had recorded deficiencies involving alarm, navigation, life-saving, and steering-related systems before its 2025 collision with the “Stena Immaculate” [13,14]. These observations further motivate the use of PSC records as pre-accident information for severity prediction.

This study therefore aims to develop a PSC-assisted maritime accident severity prediction framework that provides a basis for pre-accident assessment. In this study, prediction refers to the assessment of potential accident severity based on PSC information available before accident occurrence. Therefore, the proposed framework focuses on identifying vessel conditions associated with severe accident outcomes and serves as a predictive component supporting broader maritime accident analysis processes. To achieve this objective, an integrated analytical framework is developed to combine classical maritime accident data with PSC inspection records and introduce an advanced two-stage data augmentation strategy. The framework is designed to shift maritime safety assessment from static, retrospective analysis toward dynamic, forward-looking accident severity prediction. Specifically, PSC inspection records are incorporated as supplementary pre-accident risk information. Integrated with a global maritime accident database, these records are used to construct a multi-source prediction database for individual vessels. To overcome the challenge of limited matched samples and high-dimensional categorical features in the PSC-accident fusion database, a two-stage augmentation pipeline is designed: first, the Synthetic Minority Over-sampling Technique for Nominal features (SMOTEN) method is applied to provide foundational sample expansion, thereby meeting the basic training requirement for the subsequent diffusion-based generator; subsequently, the proposed Conditional Tabular Denoising Diffusion Probabilistic Model (CTDDPM) is utilized to generate high-fidelity synthetic samples, enabling the system to effectively learn complex risk patterns from limited matched data. By modelling and analyzing this augmented data, the system predicts potential accident severity and provides predictive support for identifying high-risk conditions for specific vessels or fleets. This approach supports a transition from post-incident analysis to pre-incident risk assessment. This study addresses four core research questions:

- 1) Matched-sample augmentation: Given the limited sample size typical of maritime accident records, how does the proposed two-stage augmentation framework enhance model performance and provide more reliable predictive support for pre-accident severity assessment?
- 2) Risk feature interpretation: Which PSC deficiency types provide the strongest predictive signals for accident severity and should therefore receive priority in feature-level interpretation?
- 3) Accident severity prediction model development: Based on the integrated PSC and accident data, which machine learning model or deep learning model can most accurately and reliably predict accident severity?
- 4) Risk-oriented decision support: How can model-derived predictive evidence, particularly the identified PSC-related features, support risk prioritization and proactive maritime safety management?

By systematically answering these questions, this study advances a pre-accident severity prediction framework with practical potential. By unlocking the predictive value of PSC data, the framework can provide supplementary evidence for risk-oriented inspection prioritization and support proactive safety-management decisions by maritime stakeholders. Ultimately, it supports the transition in maritime safety governance from a reactive, compliance-driven paradigm to a proactive, risk-based preventive approach.

The rest of this paper is organized as follows. In Section 2, a comprehensive literature review is provided, discussing the advancements in maritime accident prediction, challenges related to limited

matched samples. Section 3 presents the methodology, detailing the construction of the PSC-accident multi-source fusion framework, and the classification models used for performance evaluation. Section 4 concentrates on the feature importance analysis, particularly of PSC inspection records, while empirically validating the effectiveness of the proposed framework. Finally, Section 5 concludes the paper, highlighting the contributions of the study, its limitations, and future research directions.

2. Literature review

This section systematically reviews the scholarly foundations pertinent to developing a pre-accident severity prediction framework assisted by PSC inspection records. The review is structured into two thematic streams to address the core challenges identified: (1) the evolution of data-driven accident prediction methodologies and PSC-related maritime safety analysis, (2) approaches to limited matched samples, data scarcity, and class imbalance in maritime contexts. This structured analysis not only synthesizes the current state of knowledge but also delineates the specific research gaps that the present study seeks to address.

2.1. Advancements and data challenges in maritime accident prediction

The application of machine learning has significantly advanced maritime safety research, particularly in analysing accident causation and severity [15]. Researchers have effectively employed models like Random Forest to identify key risk factors, such as poor team communication, accident timing, and vessel type, in determining accident outcomes [16]. This predictive capability is crucial for proactive risk management [17]. Recent studies have further improved maritime accident severity prediction through feature engineering, data balancing, explainable machine learning, and multi-source feature integration, indicating that data-driven maritime accident analysis has reached a relatively mature stage. However, a fundamental constraint persists: existing maritime accident databases are often limited in feature dimensions, particularly in terms of pre-accident vessel-condition information, restricting a comprehensive analysis of risk factors and creating a bottleneck for model performance [18,19].

To overcome this data limitation, the paradigm of multi-source data fusion has emerged. Integrating external data streams, such as weather information, has been shown to improve prediction accuracy for specific accident types [4]. Most notably for this study, research has begun to link PSC inspection records with safety outcomes. Studies demonstrate that vessel attributes (age, type) and PSC inspection details significantly influence vessel risk levels and detention likelihood, validating PSC data as a potent risk indicator [18]. Other studies have further incorporated PSC deficiencies, ship risk level, PSC inspection level, and accident consequences into probabilistic risk frameworks, indicating that PSC information can support the assessment of ship accident consequences and vessel safety levels [20]. More recent studies have extended this research line through dynamic Bayesian Network modelling and causal neural network analysis to explore the dynamic or causal relationships between historical PSC deficiencies and subsequent maritime accidents [21,22]. This establishes a compelling premise: fusing PSC records, which contain rich, pre-incident deficiency information, with historical accident data can create a more powerful predictive feature set. Yet, this fusion introduces a new technical challenge: the resulting PSC-accident fusion database is characterised by high-dimensional categorical features and limited matched accident samples, which can lead to model overfitting and impaired generalization.

2.2. Mitigating data scarcity: from oversampling to generative augmentation

Data scarcity presents a significant challenge in risk prediction [23].

In maritime contexts, where serious accidents are rare, the limited amount of available data can hinder model performance, particularly in capturing rare but severe accidents [24].

Traditional methods like the Synthetic Minority Over-sampling Technique (SMOTE) and its variants operate by interpolating between existing minority-class samples. Their application in maritime studies has confirmed their utility in mitigating data imbalance and improving model accuracy [19]. However, a key limitation remains: these interpolation-based methods may generate samples that do not fully capture the complex, real-world data distribution, especially in a high-dimensional, fused feature space.

This limitation has motivated the exploration of deep-learning-based generative models. While Generative Adversarial Networks (GANs) can produce high-quality synthetic data, their training is often unstable and typically demands substantial data volumes, requirements that conflict with the limited matched-sample setting common in maritime contexts [25,26]. In contrast, Denoising Diffusion Probabilistic Models (DDPM) have recently shown superior performance in generating realistic and diverse synthetic data, even for tabular data, and often with lower sample size requirements than GANs [27,28]. This makes them a promising candidate for maritime data augmentation. Recent diffusion-based methods such as TabDDPM have extended diffusion modelling to heterogeneous tabular data [27]. However, the PSC-accident fusion dataset considered in this study is characterized by limited matched samples and predominantly encoded categorical risk indicators. This motivates a task-specific adaptation of the standard DDPM framework to provide compact conditional generation for the present data setting.

2.3. Research gaps and contributions of this study

A thorough review of the literature reveals a distinct gap between the established value of PSC data for static risk profiling and its underexplored potential for dynamic accident prediction. Existing studies have already demonstrated that PSC inspection records are associated with detention likelihood, vessel risk levels, accident involvement, and accident consequences. However, these studies have mainly focused on econometric risk estimation, explanatory Bayesian Network modelling, PSC effectiveness assessment, detention prediction, dynamic risk analysis, or causal interpretation [29]. Relatively limited attention has been paid to transforming PSC-accident fusion data into a prediction-oriented framework for pre-accident severity assessment, missing a critical opportunity for early warning. Furthermore, while multi-source data fusion is recognized as beneficial, prior studies have reported only modest performance gains and lack a systematic framework that couples advanced data augmentation with interpretable feature analysis to fully exploit the fused dataset. Simply adding more features without addressing the core issues of sample scarcity and model transparency yields suboptimal results.

Building on this analysis, this study is designed to bridge these identified gaps through three key contributions:

Contribution 1: Systematic validation of a PSC-accident fused dataset for severity prediction. This study moves beyond using Paris MoU PSC inspection records for administrative decisions and constructs a prediction-oriented multidimensional maritime safety feature set by integrating PSC inspection records with a global accident database. This study empirically validates that this fusion further enhances the performance of accident-severity prediction models, confirming PSC records' role as supplementary pre-accident indicators of accident outcomes.

Contribution 2: A task-adapted, conditional data augmentation framework. To solve the small-sample, high-dimensionality problem inherent in the fused data, a hybrid augmentation strategy is proposed. The framework first employs SMOTEN, a variant designed for categorical features, for initial sample expansion. It then develops CTDDPM as a task-adapted conditional extension of DDPM for the encoded PSC-

accident feature space. CTDDPM integrates severity-conditioned diffusion, discrete-aware regularization, and output projection into a unified generative process tailored to the limited-sample and predominantly categorical characteristics of the fused PSC-accident data.

Contribution 3: A holistic framework for enhanced prediction and interpretation. This study provides an integrated pipeline, from multi-source fusion and advanced data augmentation to Feature Selection (FS) and model interpretation. This framework does not merely aim for higher accuracy; it is designed to identify and interpret the most predictive features, particularly those from PSC records (e.g., specific deficiency types or combinations), thereby translating model predictions into interpretable predictive evidence for pre-accident oriented maritime safety management.

3. Methodology

This study integrates PSC inspection records with the IMO accident database to construct a multi-source dataset for maritime accident severity prediction. To address the dual challenges of limited matched samples and high-dimensional categorical features in the PSC-accident fused dataset, this study proposes a two-stage data augmentation and feature selection framework. As shown in Fig. 1, the first phase of basic augmentation uses SMOTEN to address matched-sample sparsity in categorical data, providing an initial sample expansion suitable for nominal features. The second phase of refined generation employs the CTDDPM, which captures complex associations across multi-source data while preserving the semantics of categorical features, thereby generating synthetic samples consistent with accident severity labels. Finally, FS is conducted based on tree model importance analysis to construct a lightweight, high-precision predictive model, with the aim of reducing feature redundancy and retaining discriminative information for downstream prediction. This hierarchical design overcomes the limitations of traditional methods in the PSC-accident fused dataset context. The FS step further ensures that the augmented feature space retains discriminative information while maintaining interpretability, thereby laying the theoretical foundation for pre-accident severity prediction.

3.1. Data acquisition

To build a comprehensive dataset for maritime accident severity modelling, this study utilizes two key sources of data: the IMO accident database and PSC inspection records. These datasets support the evaluation of how PSC inspection records contribute to predicting the severity of maritime accidents, focusing on their role in enhancing the accuracy of severity predictions.

3.1.1. IMO accident database

The accident dataset used in this study is sourced from the "Marine Safety Investigation Reports" section of the Global Integrated Shipping Information System (GISIS), administered by the IMO. These reports are submitted by IMO Member States in accordance with the Casualty Investigation Code (MSC.255(84)) and serve as one of the authoritative, standardized sources of maritime casualty data [30]. This study obtained 7941 accident records covering the period from 2000 to 2024, with each record corresponding to a single maritime safety incident. The primary information in each record is listed in Table A1 of Appendix A. The raw data encompasses multiple dimensions of information and provides contextual background for each maritime accident. A key feature of this dataset is the standardized occurrence severity classification recorded in GISIS, which categorizes cases as Marine Incident, Marine Casualty, or Very Serious Marine Casualty under the IMO's harmonized reporting procedures and the definitional framework of the Casualty Investigation Code, where "very serious" denotes a high-severity subset of casualties (e.g., total loss, fatalities, or severe environmental damage) and "incident" refers to safety-threatening occurrences that do not meet the casualty definition [30,31]. In this study,

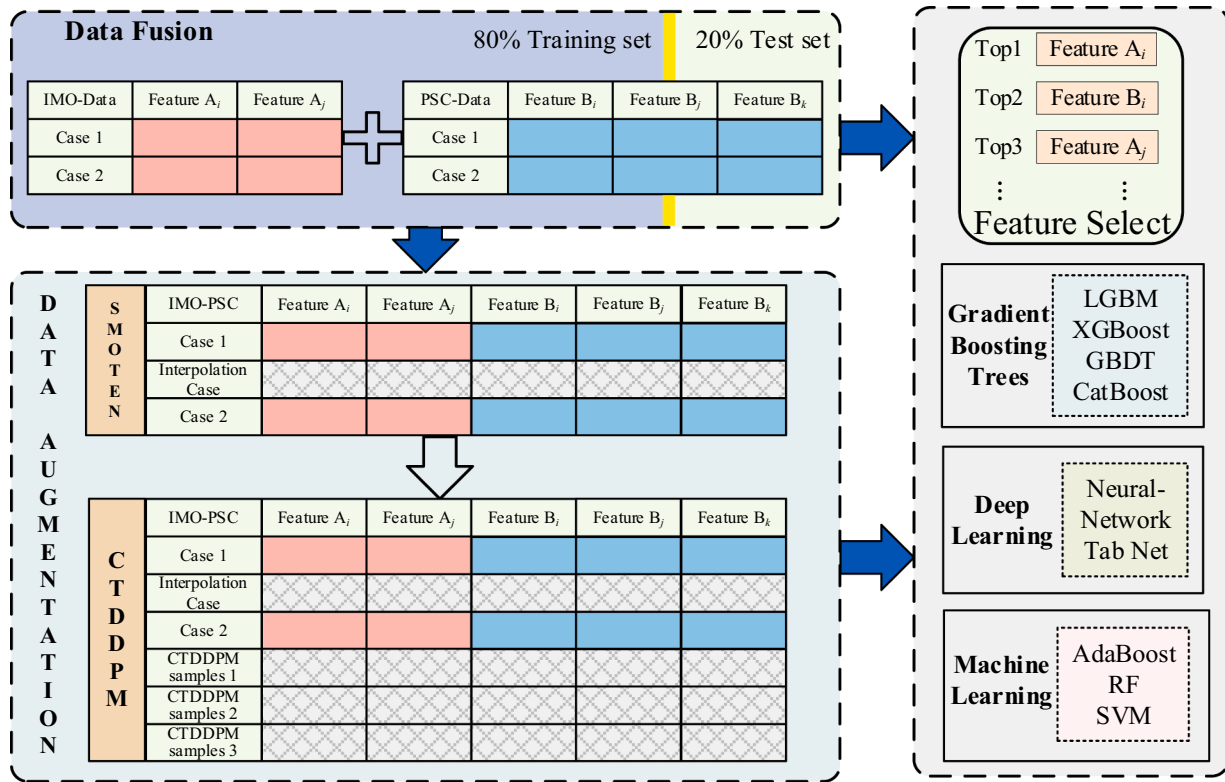


Fig. 1. The architecture of the SMOTEN+CTDDPM+FS augmentation framework.

however, the categories 'Marine Casualty' and 'Marine Incident' are both treated as less severe incidents, while 'Very Serious Marine Casualty' is categorized as a severe accident. This classification aligns with this study's focus on predicting the severity of maritime accidents, allowing for a clearer distinction between minor and major accidents. The rationale behind this classification is to emphasize the predictive modelling of serious accidents while simplifying the categorization of less severe ones.

3.1.2. PSC inspection records

PSC inspection records were downloaded in XML format from the public Data Service (DS) under the "Statistics & Current Lists" section of the official Paris MoU website. In this study, PSC inspection data were processed and categorized for improved analytical efficiency, as detailed in Table 1. The inspection types include Detailed, Expanded, and Initial inspections, reflecting the regulatory rigor applied based on vessel risk levels. Flag states are classified into White, Grey, Black, and "Other" categories, with "Other" used for flag vessels that cannot be classified due to insufficient historical inspections (i.e., fewer than 30 inspections during the statistical period). Similarly, classification societies are categorized as High, Medium, Low, Very Low, or "Unknown" based on their detention rates and compliance records. "Unknown" refers to societies that have fewer than 60 inspections, making their evaluation

uncertain. The Statutory Certificates Surveying Authority Type, based on the authority issuing statutory certificates, is categorized into Flag and RO (Recognized Organization). Vessel characteristics, including ship type, age, and gross tonnage, are categorized for subsequent modelling. Specifically, ship type and age are grouped with reference to the classification scheme in the UNCTAD's "Review of Maritime Transport 2024" [32], while gross tonnage is divided into five intervals with reference to previous PSC-related ship risk and detention studies [33, 34]. Class certificates are assessed by their issuing authority and current status, with categories like Delivered, Reassigned, Reinstated, Suspended, or Withdrawn. Detention duration is categorized into five intervals based on empirical research on detention time [34]. The deficiency codes adhere to the 2024 version of the Paris MoU Overview of Deficiency Codes, with corresponding descriptions provided in Table A2 of Appendix A, while the certificate codes are listed in Table A3 of Appendix A [11].

3.1.3. Database fusion

This study integrates PSC inspection records with the IMO accident database to effectively predict maritime accident severity. Both datasets contain critical regulatory and accident data for global vessels, with the IMO number serving as a unique vessel identifier, ensuring sample match ability across the databases. During the data integration process,

Table 1
Overview of selected features in the PSC inspection record database.

Feature	Value				
PSC Inspection Type	Initial Inspection				Expanded
Flag	White	Grey	Detailed Black	Other	
Ship Type	Oil Tankers	Bulk Carriers	General Cargo Ships	Container Ships	Other Ships
Keel Date (Years)	0-4	5-9	10-14	15-19	≥20
Gross Tonnage	<10,000	10,000-39,999	40,000-59,999	60,000-99,999	≥100,000
Class Certificates Issuing Authority	High	Medium	Low	Very Low	Unknown
Class Certificates Class Status	Delivered	Reassigned	Reinstated	Suspended	Withdrawn
Statutory Certificates Surveying Authority Type	Flag		RO		
Inspection Duration of Detention (Days)	1-3	4-7	8-14	15-28	29-84

the PSC inspection records and all vessel samples in the IMO accident database were accurately matched based on the vessel's IMO number. Some vessels involved in accidents had multiple PSC inspection records before and after the incident. To maintain the logical sequence of data and ensure predictive accuracy during model training, this study's integration strategy retains only the most recent PSC inspection record prior to the accident as the representative regulatory feature for each vessel. The 36-month lookback window is consistent with the Paris MoU risk-assessment framework, in which the Ship Risk Profile incorporates a 36-month inspection history and company performance, while flag performance is evaluated over a three-year rolling period. Accordingly, to preserve temporal relevance, PSC inspections conducted >36 months before the accident are excluded. After the matching and temporal filtering procedures, 1125 accident samples were successfully matched with eligible PSC inspection records from the original 7941 GISIS accident records, corresponding to a match rate of 14.17%. The matched dataset consisted of 523 Class 0 samples (46.49%) and 602 Class 1 samples (53.51%). Based on the vessel's IMO number, this study integrates the accident vessel's historical regulatory information, providing rich regulatory features and reliable accident labels for model building. As this study focuses on evaluating the contribution of PSC inspection records in predicting accident severity, the features from these records were categorized for convenience in further analysis. The categorization of PSC inspection features, as shown in Table 2, facilitates the process of feature engineering and ensures the model's scalability in subsequent analyses.

3.2. Two-Stage data augmentation method

To tackle the generative modelling challenges of the PSC-accident fused dataset under limited sample availability and categorical features, this study develops a two-stage data augmentation framework aimed at improving both the quantity and quality of training data. Before data augmentation, the original fused dataset was divided into training and test sets. All data augmentation and generator training procedures were performed exclusively on the training set. For all four generative models (CTDDPM, TabDDPM, CTGAN, and TVAE), the model hyperparameters were specified independently of the holdout test data, and no test-set samples or labels were used for generator training or hyperparameter selection. The holdout test set was excluded from model fitting and synthetic-data generation and was used to assess downstream predictive performance.

3.2.1. First-stage augmentation: SMOTEN for imbalanced data

Although diffusion models are generally less demanding in terms of data volume, achieving high-quality generation in complex tasks and multi-source data fusion settings typically still requires substantial training data. In addition, SMOTE has proven effective for oversampling minority classes in imbalanced datasets and can broaden the sample space under small-sample conditions [19]. Therefore, SMOTE has the

potential to serve as a basic data augmentation method, providing foundational expansion in the sample dimension. Originally designed for numerical features, SMOTE generates new samples through linear interpolation of minority class samples to achieve class balance. However, the data in this study's PSC-accident fused dataset are categorical, making traditional SMOTE unsuitable, as direct interpolation would disrupt the discrete nature of the variables. Synthetic Minority Over-sampling Technique for Nominal features (SMOTEN) is an extension of the classic SMOTE algorithm, specifically designed for categorical data. It selects neighbouring samples from the nearest minority class samples and randomly replaces categorical attributes based on consistency, rather than performing numerical interpolation. Therefore, it is suitable for use with the PSC-accident fused dataset in this study. By using SMOTEN as the first-phase augmentation method, this approach provides basic expansion in the sample dimension, offering initial support for generating more complex diffusion models in subsequent stages. This strategy enhances the model's ability to learn categorical features and provides more representative input samples for deep generative models, addressing the issue of insufficient model training in small sample scenarios.

3.2.2. Second-stage augmentation: conditional tabular DDPM

(1) The traditional DDPM.

Traditional DDPM is an unconditional generative model that has successfully generated continuous data, such as images. The process of generating samples with DDPM involves two steps.

Step 1: The model starts with real data (e.g., images or tabular data) and gradually adds noise until a pure noise sample is obtained. The forward diffusion process, starting with the original data x_0 , is represented by the joint distribution in Eq. (1). Here, $q(x_t|x_{t-1})$ represents the conditional probability distribution from x_{t-1} to x_t , modelled as a Gaussian distribution in Eq. (2). The conditional distribution $q(x_t|x_{t-1})$ is applied to the original data x_0 , generating a series of noisy samples, $x_1, x_2, \dots, x_T$. By adjusting the β_t values, the model gradually introduces noise until the data becomes pure Gaussian noise. The goal is to model the entire forward diffusion process during training, learning the distribution of the noisy data.

$$q(x_1, \dots, x_T|x_0) = \prod_{t=1}^T q(x_t|x_{t-1}) \quad (1)$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \quad (2)$$

where T represents the total number of timesteps in the diffusion process. β_t is a variance schedule parameter at timestep t . I denotes the identity matrix.

Step 2: The reverse process is achieved through the Gaussian transition process Eq. (3). Starting from the noisy sample x_T , the model gradually removes noise and recovers the original data x_0 through the trained noise prediction network. At each time step t , the denoising process is described by the conditional probability distribution $p_\theta(x_{t-1}|x_t)$, which is learned during training. The goal is to progressively recover the data distribution and ultimately reconstruct the original data x_0 from the noisy samples. The conditional probability $p_\theta(x_{t-1}|x_t)$ is modelled as a Gaussian distribution, as shown in Eq. (4)

$$p_\theta(x_0, \dots, x_{T-1}|x_T) = \prod_{t=1}^T p_\theta(x_{t-1}|x_t) \quad (3)$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta^2(x_t, t)I) \quad (4)$$

where $\mu_\theta(x_t, t)$ represents the predicted mean, indicating the denoised output at time step t , $\sigma_\theta^2(x_t, t)$ is the predicted variance, controlling the noise level during the denoising process, and I is the identity matrix.

(2) The CTDDPM.

Table 2

Hierarchical structure of feature labels in the PSC inspection records.

Level 1 Label	Level 2 Label	Level 3 Label
PSC Inspection Records	Deficiencies	Deficiencies Code
	Certification	Certification Documents Code
	Documents	
	PSC Inspection Information	PSC Inspection Type
		Flag
		Class Certificates Issuing Authority
		Class Certificates Class Status
		Statutory Certificates Surveying Authority Type
		Inspection Duration of Detention
	Ship Particulars	Ship Type
		Keel Date
		Gross Tonnage

The introduction of PSC data in the PSC-accident fused dataset significantly increases the feature dimensions, amplifying the negative impact of insufficient sample size in the original data and generating more complex relationships between features. However, the standard DDPM architecture is primarily designed for generating continuous data, such as pixel values in images, and has limited capability in handling categorical labels. It is generally unsuitable for processing complex tabular data, especially when categorical and high-dimensional features are involved. As a result, the original DDPM model performs poorly when handling PSC-accident data. To address this issue effectively, this study proposes the Conditional Tabular DDPM, an extension of DDPM.

CTDDPM is developed by adapting the standard DDPM framework to the characteristics of the PSC-accident fusion data. Conventional image-oriented DDPM implementations commonly employ U-Net-based denoisers to exploit spatial dependencies in image data. In contrast, the PSC-accident samples are represented as fixed-length encoded feature vectors without such spatial topology. CTDDPM therefore adopts an MLP-based noise-prediction network in place of the image-oriented denoising architecture. To enable targeted augmentation, the diffusion timestep and accident-severity label are represented through learnable embeddings and concatenated with the noisy feature vector before being passed to the denoising network.

In the implementation, CTDDPM used 1500 diffusion timesteps with a squared-cosine noise schedule. The denoising MLP contained two hidden layers with 64 units each and SiLU activation, with a dropout rate of 0.1. The timestep and class-label embedding dimensions were set to 8 and 4, respectively. The model was trained for 500 epochs using AdamW with a learning rate of 5×10^{-4} and a batch size of 8. The weight of the auxiliary discrete-aware regularization term was set to 1.0.

To accommodate the discrete characteristics of the encoded PSC features, an auxiliary discrete-aware regularization term is incorporated into the noise-prediction objective, and the outputs obtained after the reverse diffusion process are projected onto predefined discrete values. In this way, CTDDPM retains the Gaussian forward and reverse diffusion mechanism of DDPM while adapting the denoising architecture, conditioning mechanism, and discrete-data processing to the PSC-accident setting.

CTDDPM is designed to address the specific characteristics of the encoded PSC-accident feature space. TabDDPM employs a hybrid formulation combining Gaussian diffusion for numerical variables and multinomial diffusion for categorical variables [27]. CTDDPM adopts a unified severity-conditioned Gaussian diffusion process over the encoded feature vectors, integrating discrete-aware regularization and output projection to accommodate their categorical characteristics. These components collectively establish a task-adapted generation procedure for the limited-sample PSC-accident dataset. The architectural and methodological characteristics of DDPM, TabDDPM, and CTDDPM are summarized in Table 3.

Based on these adaptations, CTDDPM enhances the traditional DDPM model by combining label embedding and timestep embedding, enabling effective generation of conditional tabular data. The CTDDPM process is illustrated in Fig. 2, and the corresponding end-to-end training and sampling procedure is summarized in Algorithm B1 of Appendix B. First, label embedding maps categorical labels into a lower-dimensional

vector space. Then, as shown in the pseudocode in Algorithm 1, label embedding and timestep embedding are input into the model, along with other feature data. This enables the model to not only determine which category each sample belongs to, but also to understand the intensity of the current noise. Simultaneously, label embedding acts as a conditional input, enabling the model to generate data that conforms to the specified labels. Finally, as shown in the pseudocode in Algorithm 2, the difference between the generated sample and the discrete label is calculated, mapping continuous values to the closest discrete labels, thereby ensuring the generated samples meet the requirements of the discrete category labels. It is worth noting that, to accommodate categorical tabular inputs, this study adopts a Multi-Layer Perceptron (MLP) based denoiser in place of the conventional U-Net, which is primarily tailored to spatially structured data.

(3) The similarity comparison between real and synthetic data.

To evaluate the quality of categorical data, this study measures the distribution differences of each categorical variable using Total Variation Distance (TVD) and constructs the Column Shapes Score, reflecting the overall similarity between synthetic and real data in univariate distributions. For a categorical variable X_i , the TVD calculation process for its distribution difference between real and synthetic data is shown in Eq. (5). Based on Eq. (5), the corresponding TVD value is calculated for each column, and the mean of the TVD values across all columns is used to obtain the global univariate distribution deviation. To make this metric more intuitive, it is converted into a similarity score, the Column Shapes Score, using Eq. (6). The Column Shapes Score ranges from 0 to 1, with a higher score indicating that the synthetic data is closer to the real data at the single-column feature level.

At the multivariate level, this study calculates the correlation differences between all variable pairs in real and synthetic data using Cramér's V and transforms them into the Column Pair Trends Score. First, for each pair of categorical variables (X_i, X_j) in the real and synthetic data, a contingency table is constructed, and the Cramér's V value is calculated according to Eq. (7). Then, the absolute difference in Cramér's V ΔV_{ij} for the same variable pairs in the real and synthetic data is calculated according to Eq. (8). Finally, as shown in Eq. (9), the Column Pair Trends Score is obtained by subtracting the Mean Absolute Error (MAE) of all variable pairs from 1.

$$TVD(P, Q) = \frac{1}{2} \sum_{x \in K} |P(x) - Q(x)| \quad (5)$$

$$Column\ Shapes\ Score = 1 - \frac{1}{d} \sum_{i=1}^d TVD(P_{X_i}, Q_{X_i}) \quad (6)$$

$$V_{ij} = \sqrt{\frac{\chi^2/n}{\min(k_i - 1, k_j - 1)}} \quad (7)$$

$$\Delta V_{ij} = |V_{ij}^r - V_{ij}^s| \quad (8)$$

$$ColumnPairTrendsScore = 1 - \frac{1}{N} \sum_{i < j} \Delta V_{ij} \quad (9)$$

where X_i represents the i^{th} categorical variable in the dataset. K denotes

Table 3
Comparison among original DDPM, TabDDPM, and CTDDPM.

Aspect	Original DDPM	TabDDPM	CTDDPM
Feature representation	Continuous-data representation	Numerical and one-hot categorical features	Directly encoded fixed-length PSC-accident feature vectors
Denoising network	U-Net	MLP	MLP
Diffusion formulation	Gaussian diffusion	Gaussian diffusion for numerical variables and multinomial diffusion for categorical variables	Unified conditional Gaussian diffusion with a DDPM scheduler for task adaptation
Conditioning integration	No class condition	Timestep and class information incorporated into the MLP	Direct concatenation of noisy features, timestep embedding, and accident-severity label embedding into the MLP

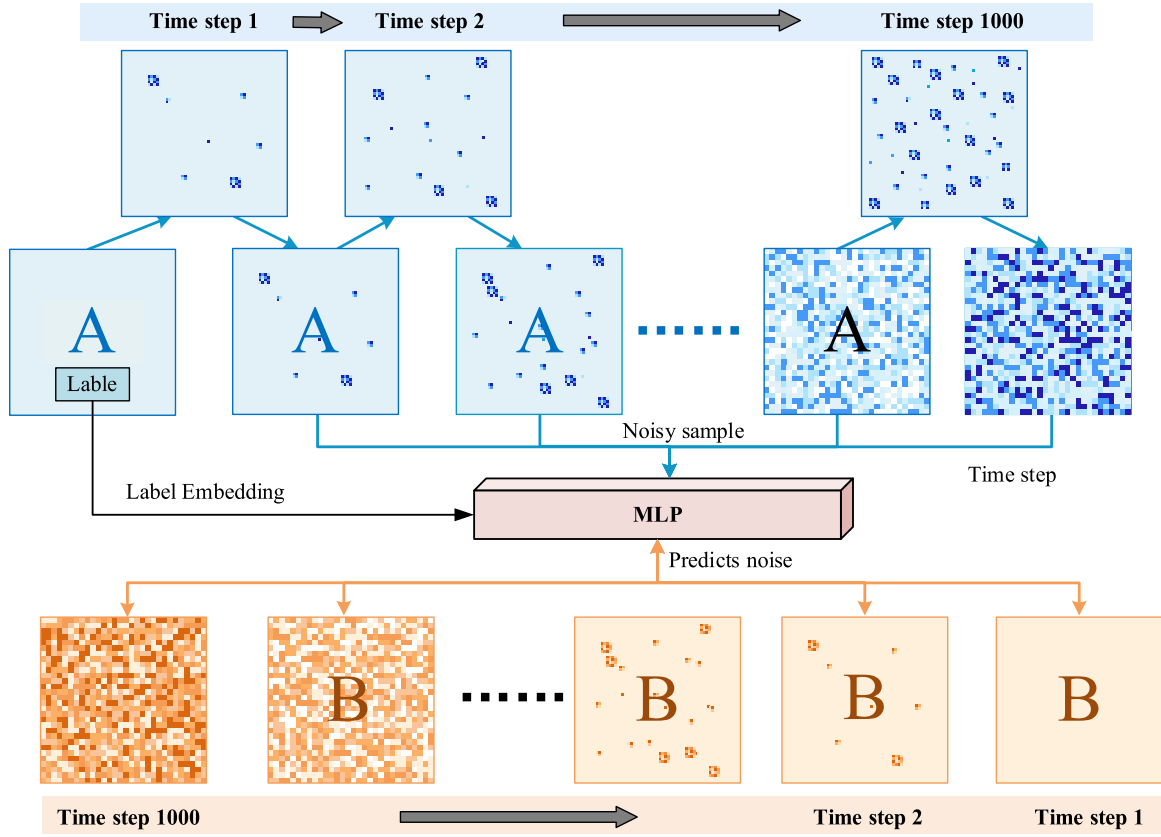


Fig. 2. Forward and reverse processes of CTDDPM.

Algorithm 1

Label embedding and time step embedding for ConditionalTabDDPM.

Input:	Labels; t : Time step; num_labels; embedding_dim: Dimensionality of the embedding space; num_timesteps: Total number of time steps
Output:	Generated data based on labels and time step information
1	Begin
2	Initialize label embedding layer
3	Initialize the time step embedding layer
4	For each input label y :
5	map the label to its embedding vector
6	For each input time step t :
7	map the time step to its embedding vector
8	Concatenate the label embedding y_{embed} and time step embedding t_{embed} with other feature data x
9	Pass the combined input through the model to generate the output data
10	Return the generated output
11	End

Algorithm 2

Discretizing generated samples for CTDDPM.

Input:	sample: Generated continuous samples (from the model) discrete_values: Predefined set of discrete values (e.g., class labels)
Output:	discrete_samples: Discretized samples (mapped to discrete values)
1	Begin
2	Calculate the absolute difference between the generated sample and each discrete value
3	Find the closest discrete value for each generated sample
4	Map each generated sample to its closest discrete value
5	Return the discretized samples
6	End

the set of all possible categories that X_i can take. $P(x)$ is the frequency of category x in the real data. $Q(x)$ is the frequency of category x in the synthetic data. d is the total number of categorical variables in the

dataset. χ^2 is the chi-square statistic derived from the contingency table of variables X_i and X_j , n is the number of samples, and k_i and k_j denote the number of unique categories in variables X_i and X_j , respectively. V_{ij}^r and V_{ij}^s are the Cramér's V values for variables X_i and X_j in the real and synthetic datasets, respectively. N is the total number of unique variable pairs.

3.3. Tree-based importance analysis

To address the high-dimensional categorical data in the PSC-accident fused dataset, this study introduces Forward Feature Selection (FFS) based on importance ranking during the feature engineering phase. This method combines the embedded feature importance capability of ensemble tree models with the stepwise validation mechanism of a forward search strategy, providing both data-driven insights and model adaptability. Specifically, using the training set, an initial model is built with basic machine learning models, and its feature importance scores, based on split counts, are extracted to rank all input features. Then, starting with the highest-scoring feature, features are introduced sequentially according to the ranking. After each new subset is added, the model is retrained, and its performance metrics are evaluated. This process continues until all features are incorporated.

Compared to traditional filter-based or wrapper-based feature selection methods, FFS is more suitable for the PSC-accident fused dataset. On one hand, the features in PSC inspection data and IMO accident records are all categorical variables, and tree-based feature importance assessment naturally handles categorical data without requiring manual encoding or additional feature engineering. On the other hand, the PSC-accident fused dataset has high-dimensional features and complex inter-feature correlations. Simple univariate filtering methods may fail to capture the nonlinear effects of feature combinations, whereas this method, based on the global feature contributions learned by the model,

more effectively identifies feature subsets with actual predictive value. Furthermore, the stepwise introduction and real-time evaluation of performance changes make the inclusion of each feature verifiable, significantly reducing the risk of overfitting due to redundant features. This is particularly suitable for small sample, categorical-dominant, and feature-rich complex fusion data scenarios.

In summary, the FFS method used in this study enhances the controllability and transparency of model training while providing a lightweight, highly interpretable, and generalizable input feature set for the final accident-severity prediction model.

3.4. Predictive modelling and evaluation

To assess the effectiveness of the proposed data augmentation method, this study evaluates the performance of multiple classification models. By selecting a diverse set of models that represent different modelling approaches, this analysis provides insights into how well these methods handle the challenges posed by high-dimensional, small-sample maritime accident data.

3.4.1. Predictive modelling

To comprehensively assess the applicability and performance impact of the proposed data augmentation method across various classification models, this study selects nine widely used and representative models for the experimental analysis. These models include gradient boosting ensemble methods like LGBM, XGBoost, GBDT, CatBoost, and AdaBoost; traditional ensemble models such as RF; deep learning models like TabNet and Neural Networks; and classical machine learning models such as SVM. The selected models cover the main modelling mechanisms most suitable for static categorical tabular classification under limited matched-sample conditions, including boosting-based ensemble learning, bagging-based ensemble learning, kernel-based learning, and neural-network-based tabular learning. Each of these models represents a distinct modelling mechanism, with a solid theoretical foundation, mature engineering implementations, and a wide range of practical applications. They fully capture the response characteristics of various models to limited matched-sample, high-dimensional categorical datasets. Table 4 provides a comprehensive summary of the main features, advantages, and applicable scenarios of the nine classification models selected in this study.

3.4.2. Error-based evaluation of probabilistic outputs

Traditional classification metrics, while commonly used for performance evaluation, may not fully capture the complexities inherent in probabilistic models, especially when dealing with uncertain and high-dimensional data. This study employs an error evaluation method

based on the probabilistic outputs of classification models to systematically assess their performance in predicting accident severity. By comparing predicted probabilities for the positive class against actual binary labels, the approach quantifies the models' ability to characterize gradations of risk. On one hand, maritime accident severity involves inherent uncertainty. Even for the same type of event, its consequences may exhibit probabilistic characteristics due to variations in vessel conditions, emergency response, environmental factors, and other variables [44]. On the other hand, after integrating PSC inspection record data, the feature space becomes high-dimensional. When models encounter challenges such as sparse samples and high-dimensional variables, the credibility and rationality of output probabilities are crucial for performance evaluation [45]. Furthermore, relevant studies indicate that probabilistic error metrics such as MSE and Log Loss comprehensively reflect classification model performance in terms of confidence expression, distribution rationality, and ranking quality. These metrics are particularly suitable for complex tasks like uncertainty modelling and risk ranking, and are widely used in predictive modelling for safety-critical fields [46].

To comprehensively evaluate the probabilistic prediction capability of the models, this study selects four typical error metrics, as shown in Table 5: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Log Loss. MSE measures the squared distance between predicted probabilities and true labels, emphasizing overall calibration; RMSE, as its square root form, makes the magnitude of error more intuitively understandable; MAE focuses on average deviation and is more robust; and Log Loss, as cross-entropy loss, is the standard metric for measuring prediction confidence and distribution fitting in deep models. It heavily penalizes high-confidence incorrect predictions, making it especially suitable for handling boundary ambiguity in multi-source augmented samples in this study.

Table 5

Error metrics used for probabilistic model evaluation.

Metric	Formula	Insight Provided
MSE	$\frac{1}{n} \sum_{i=1}^n (\hat{p}_i - y_i)^2$	Captures overall calibration quality and average squared error
RMSE	$\sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{p}_i - y_i)^2}$	Reflects prediction dispersion; easier to interpret numerically
MAE	$\frac{1}{n} \sum_{i=1}^n \hat{p}_i - y_i $	Quantifies the average deviation between predicted probabilities and the true binary labels (0 or 1).
LogLoss	$-\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)]$	Assesses confidence calibration and penalizes incorrect certainty

Table 4

Characteristics and advantages of selected predictive models.

Model Name	Type	Key Characteristics	Suitable Scenarios	References
LGBM	Gradient Boosted Decision Tree	Leaf-wise growth, histogram-based splitting, fast training, strong handling of large-scale data	Large datasets, high-dimensional, categorical-rich features	[35]
XGBoost	Gradient Boosted Decision Tree	Built-in regularization, efficient missing value handling, robust generalization	Complex datasets with strong non-linear relationships	[36]
GBDT	Original Gradient Boosted Model	Additive model with stable training, effective for moderate complexity problems	Medium to small datasets, tasks requiring careful feature engineering	[37]
CatBoost	Gradient Boosted Decision Tree	Native categorical feature support, reduced overfitting tendency, automated training	Categorical feature-dominant datasets, moderate data volumes	[38]
AdaBoost	Ensemble of Weak Learners	Iteratively focuses on hard-to-classify samples, sensitive to noise, fast convergence	Low-noise environments, quick modeling with small datasets	[39]
RF	Bagging-based Ensemble Model	Strong diversity among trees, robust to overfitting, fast training, sensitive to sparse high-dimensional features	Small to medium datasets, tasks prioritizing stability	[40]
Tab Net	Deep Neural Network (Tabular Data Specialized)	End-to-end learning with feature selection, interpretable, high computational cost	Large-scale tabular datasets, requiring end-to-end deep learning	[41]
Neural Network	Standard Multilayer Perceptron	Powerful non-linear modeling capacity, prone to overfitting on small data	Large datasets with complex non-linear feature interactions	[42]
SVM	Support Vector Machine	Excels on small and sparse datasets, margin maximization, sensitive to data volume	Small-scale, sparse, or clearly separable classification problems	[43]

In Table 5, n denotes the number of instances; $\hat{p}_i \in [0, 1]$ is the predicted probability for instance i , and $y_i \in \{0, 1\}$ is the corresponding ground-truth label.

4. Results and discussion

This section examines whether the proposed PSC-accident multi-source fusion framework improves accident severity prediction and whether the resulting model outputs can generate interpretable safety insights. The analysis is organised around four linked questions: whether the synthetic data preserve useful distributional information and improve predictive performance, which PSC-related features contribute most to severity assessment, which modelling approaches provide reliable predictive support, and how the resulting predictive evidence can inform risk prioritization and proactive maritime safety management. By combining predictive evaluation, feature-level interpretation, and ablation analysis, the results provide both methodological evidence for the proposed augmentation framework and practical implications for PSC inspection prioritisation and proactive maritime safety management.

4.1. Data augmentation and predictive performance evaluation

This study systematically assesses the effectiveness of CTDDPM in data augmentation by investigating the impact of the number of generated samples on the performance of downstream prediction models. The CTDDPM model was used to generate 1000, 2000, and 3000 augmented samples, which were then integrated with the original training set to train multiple machine learning classification models, including traditional machine learning models and deep learning models (such as TabNet and Neural Networks). The target-class distributions of the datasets before and after augmentation are summarized in Table 6.

Before comparing predictive performance under different augmentation settings, the final configurations of the nine prediction models were fixed to ensure consistency across experiments. The key hyperparameter settings used in the final experiments are provided in Table D1 of Appendix D. A random seed of 1 was used for the original single-run experiments where supported by the corresponding implementation. Additional evaluations across five random seeds were subsequently conducted to assess the robustness of the principal experimental comparisons, as detailed in Sections 4.1 and 4.4 and Appendices C and D. The prediction models were implemented using Python 3.10.13. The main machine-learning libraries included scikit-learn 1.5.2 and PyTorch 1.13.1+cpu. The hyperparameters of the prediction models were determined based on the characteristics of the dataset, including the sample size and feature dimensionality, together with commonly adopted configurations and practical implementation experience. Since the objective of this study is to evaluate the contribution of the proposed data augmentation and feature-selection framework rather than to maximize the performance of an individual classifier

through extensive tuning, the same fixed model configurations were maintained across all comparative experiments. In addition, SVM inputs were standardized using StandardScaler, while the Neural Network and TabNet models were trained using the architecture- and training-specific settings reported in Table D1 of Appendix D.

As shown in Fig. 3, the use of augmented samples consistently reduces prediction error, with all models benefiting from the integration of augmented samples compared to the original dataset. Furthermore, the model prediction error decreases most significantly when the number of generated samples reaches 2000. For example, the MSE of the LGBM model decreased by 1.69% compared to when only 1000 samples were integrated, and most models showed optimal performance under these conditions. However, when the number of augmented samples increased to 3000, model performance generally declined. This phenomenon can be attributed to two interwoven theoretical mechanisms. On one hand, excessive synthetic samples, while providing diverse information, introduce additional noise, weakening the model's fit to the real data distribution and reducing the signal-to-noise ratio of training [47]. On the other hand, as the number of generated samples increases, the CTDDPM model may be prone to diversity saturation, where generated data tend to concentrate or repeat in the feature space. The information gain from new samples approaches zero and may even negatively affect the model's generalization ability. Similar diversity-related limitations have been reported in GANs and other generative models [48–50].

Based on the downstream results in Fig. 3, the 2000-sample setting was retained as the common augmentation scale for the subsequent comparison of generative methods. In addition to predictive performance, the distributional characteristics of the generated samples were evaluated using the Column Shapes Score and Column Pair Trends Score defined in Section 3.2.2 through Eqs. (6) and (9), which quantify univariate distribution similarity and pairwise relationship consistency, respectively. The detailed distributional-fidelity results for CTDDPM, CTGAN, TVAE, and TabDDPM are reported in Table D2 of Appendix D. A complementary t-SNE visualization comparing real samples with those generated by CTGAN and CTDDPM is presented in Fig. D1 of Appendix D. Under the same 2000-sample augmentation scale, the four generative methods were subsequently evaluated using the same nine downstream prediction models to examine their contribution to accident-severity prediction.

As shown in Fig. 4, CTDDPM achieved the lowest average MSE across the nine prediction models (0.2368), followed by CTGAN (0.2391), TabDDPM (0.2413), and TVAE (0.2494). CTDDPM also yielded the lowest MSE for five of the nine individual prediction models. In particular, its average MSE was approximately 1.86% lower than that obtained with TabDDPM, providing direct comparative evidence for the task-adapted conditional diffusion design under the present PSC-accident prediction setting. The complete model-specific MSE results are provided in Table D3 of Appendix D. The robustness of the CTDDPM–TabDDPM comparison was further evaluated across five matched random-seed runs (seeds 1–5), using identical holdout partitions and downstream evaluation settings within each paired run. CTDDPM achieved lower average MSE in all five runs, with an overall MSE of 0.232471 ± 0.003817 , compared with 0.239360 ± 0.001596 for TabDDPM. This corresponds to an average relative MSE reduction of approximately 2.88%, with a statistically significant difference confirmed by a two-sided paired t -test ($p = 0.0182$). The detailed results are presented in Table D4 of Appendix D. To further evaluate the robustness of CTDDPM, a one-factor-at-a-time sensitivity analysis was conducted on three key hyperparameters: the number of diffusion timesteps, the discrete-aware regularization weight (λ), and the denoiser hidden dimension. The best-performing downstream prediction model identified in the main experiments was kept fixed across all configurations under the same holdout partition. As shown in Table D5 of Appendix D, the relative changes in downstream MSE remained within approximately 3% of the predefined configuration across the examined parameter settings, indicating reasonable robustness to these

Table 6
Sample size and target-class distribution before and after data augmentation.

Dataset	Total samples	Class 0, n (%)	Class 1, n (%)
Before augmentation	1125	523 (46.49%)	602 (53.51%)
Training set before augmentation	900	408 (45.33%)	492 (54.67%)
Holdout test set	225	115 (51.11%)	110 (48.89%)
After adding 1000 synthetic samples	1900	908 (47.79%)	992 (52.21%)
After adding 2000 synthetic samples	2900	1408 (48.55%)	1492 (51.45%)
After adding 3000 synthetic samples	3900	1908 (48.92%)	1992 (51.08%)

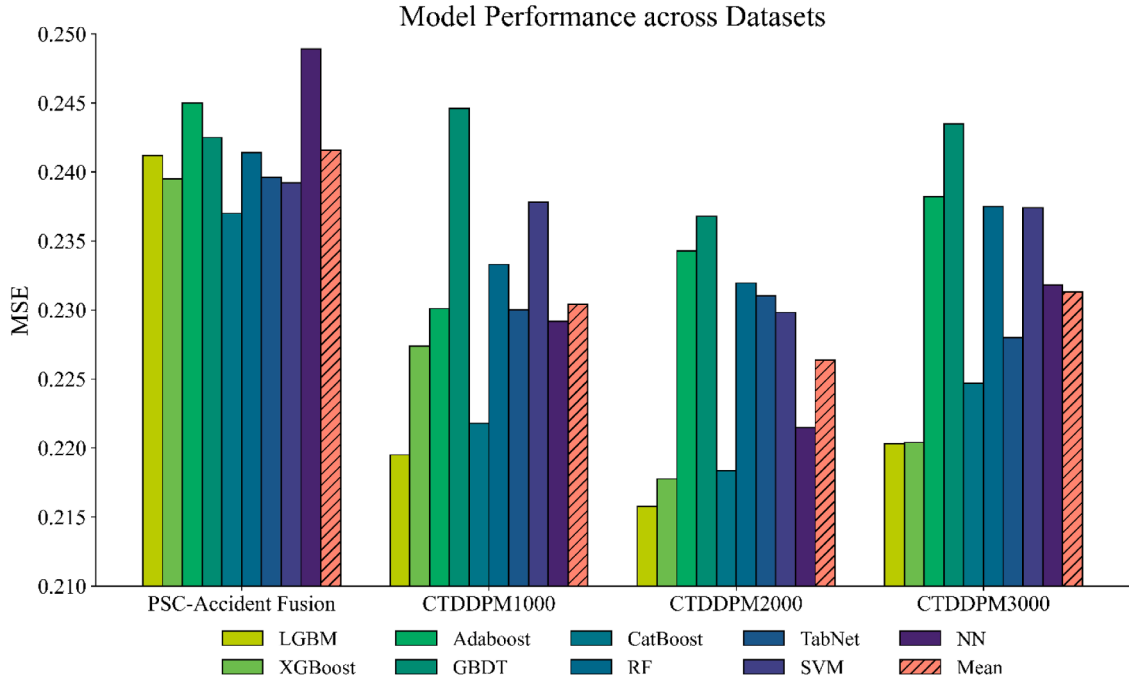


Fig. 3. The MSE results of CTDDPM under different sample sizes and comparison with original data.

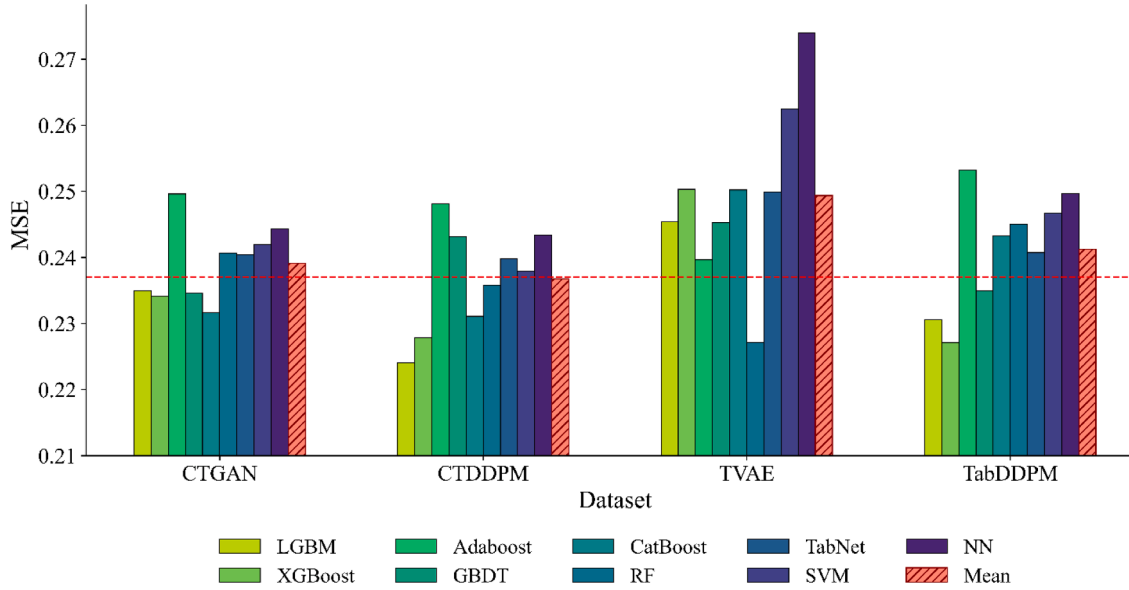


Fig. 4. Model performance across generative methods at the 2000-sample baseline.

hyperparameter variations.

Although the GAN- and VAE-based models generated samples with higher distributional fidelity, CTDDPM achieved the lowest average downstream prediction error. This difference may partly arise because high distributional similarity primarily reflects faithful reproduction of the observed training distribution, while effective data augmentation may also benefit from useful variation beyond information already represented by the limited original samples. Previous deep-learning studies have shown that moderate noise injection can act as a regularizing perturbation and improve optimization, robustness, and adaptability [51–54]. The stochastic perturbation and denoising mechanism underlying diffusion models is consistent with this principle: moderate variation around the observed feature space may provide additional training information without requiring synthetic samples to maximize

point-by-point distributional similarity. Accordingly, for the present framework, the practical value of synthetic augmentation is reflected primarily in its contribution to downstream accident-severity prediction, while the distributional-fidelity metrics provide complementary evidence regarding the statistical characteristics of the generated samples.

In summary, the results show that the effectiveness of synthetic augmentation depends not only on the generation scale but also on whether the additional samples provide useful information for downstream prediction. The 2000-sample setting produced the most favourable overall predictive performance for CTDDPM, while the cross-generator comparison showed that CTDDPM achieved the lowest average MSE despite CTGAN and TVAE obtaining higher distributional-fidelity scores. These findings support the joint consideration of

distributional fidelity and task-oriented predictive utility, with downstream accident-severity prediction serving as the principal performance criterion for the present augmentation framework.

4.2. Risk factors identification via feature importance

The feature-analysis stage of the proposed framework is primarily designed to identify a compact feature subset that directly optimizes downstream predictive performance. Since this requires not only ranking candidate variables but also determining the specific feature subset and feature number associated with the best model performance, tree-based importance ranking combined with forward feature selection is adopted as the core feature-optimization procedure in this study. After the optimal feature subset is determined, an additional SHAP analysis is conducted on the corresponding LGBM model to comparatively examine the predictive contributions of the selected features and the consistency of the resulting feature patterns.

4.2.1. Tree-based ranking and forward feature selection

This study aims to assess the contribution of various features from the integrated PSC inspection records and the IMO accident database in predicting the severity of maritime accidents. Because the proposed framework requires not only an importance ranking of candidate features, but also an explicit determination of the feature subset and feature number that yield the best downstream predictive performance, tree-based ranking combined with forward feature selection is adopted for feature optimization in this study. The analysis results, shown in Fig. 5 and Table 7, present a bar chart ranking features from high to low based on their evaluated importance, with different colours representing feature categories. The line chart on the right y-axis shows the trend in model MSE as features are sequentially introduced based on importance. As more high-value features are introduced, MSE decreases, reaching its lowest value at the final feature, indicating that the selected subset captures the main predictive information. The pie chart in the upper-

right corner summarizes the proportion of each feature category in the overall importance. The accident records and ship particulars categories account for 32.6% and 28.6%, respectively, playing a central role in predicting accident severity. Although the PSC Inspection Information, Deficiencies, and Certification Documents categories each account for a smaller proportion, they collectively provide indispensable supplementary dimensions, enriching the model's perception and enhancing its generalization ability.

The average feature importance analysis shows that the features in the Accident Records category have the highest average importance. This distribution structure reveals that features in the Accident Records category play a dominant role in predicting the severity of maritime accidents, highlighting their key role in risk modelling. The importance scores of categories such as Ship Particulars and PSC Inspection Information are also among the highest, indicating that regulatory inspection records and static vessel attributes have stable explanatory power and predictive value in prediction modelling. It is noteworthy that features in the Certification Documents category also show significant importance in the model's prediction of accident severity, indicating that the presence or absence of certain types of certificates reflects the vessel's category and potential risk characteristics to some extent. These certificates are not the direct cause of accidents, but due to the vessel types or operational characteristics they are targeted at, they are indirectly related to changes in accident severity. Deficiencies, as a micro representation of vessel safety status, usually point to specific structural or operational issues. They are important clues for identifying systematic fault chains and help in early detection of accident signs. These results validate the feasibility and effectiveness of introducing heterogeneous regulatory data such as PSC inspection information, highlighting the advantages of multi-source data fusion strategies in representing the multidimensional structure of maritime risks. This provides strong empirical support for building a systematic, data-driven maritime risk assessment framework.

Given the study's focus on PSC-assisted severity assessment, the

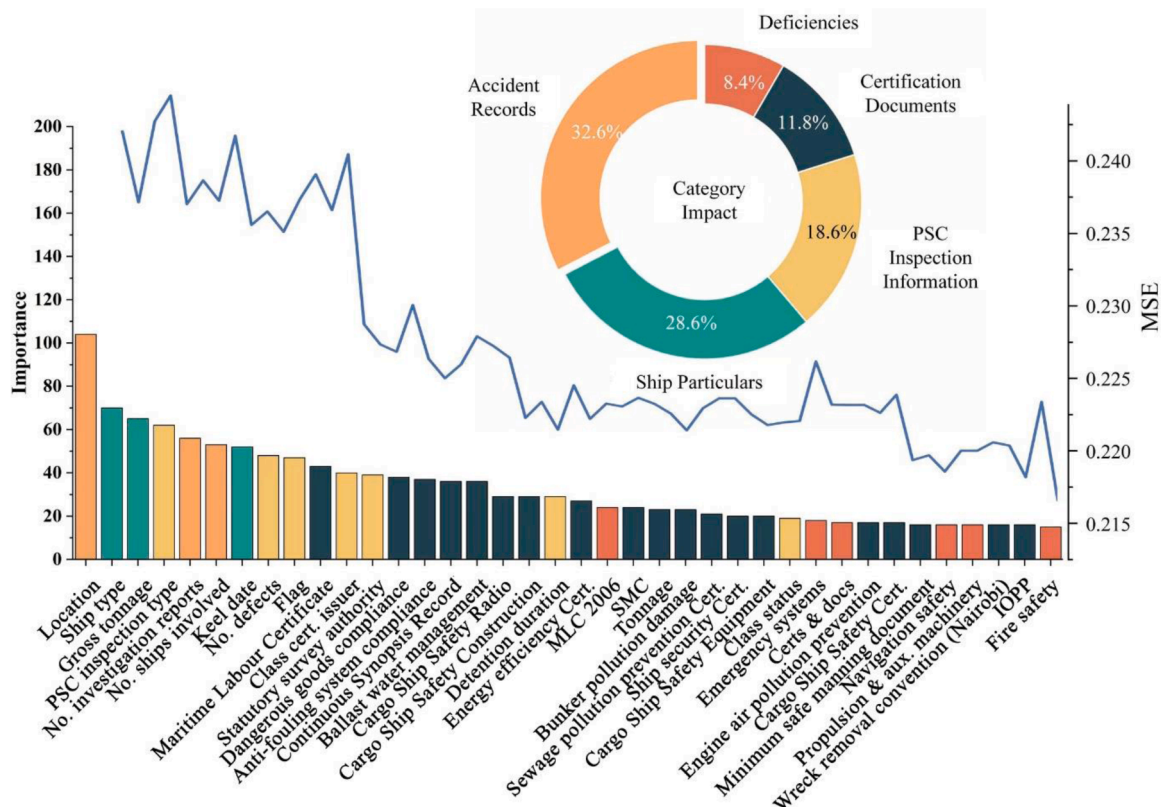


Fig. 5. Feature importance ranking and cumulative model MSE reduction.

Table 7

Overall ranking and importance scores of input features.

Rank	Feature	Importance	Rank	Feature	Importance
1	Location	104	20	Energy efficiency Cert.	27
2	Ship type	70	21	MLC 2006	24
3	Gross tonnage	65	22	SMC	24
4	PSC inspection type	62	23	Tonnage	23
5	No. investigation reports	56	24	Bunker pollution damage	23
6	No. ships involved	53	25	Sewage pollution prevention Cert.	21
7	Keel date	52	26	Ship security Cert.	20
8	Number of deficiencies	48	27	Cargo Ship Safety Equipment	20
9	Flag	47	28	Class status	19
10	Maritime Labour Certificate	43	29	Emergency systems	18
11	Class cert. issuer	40	30	Certs & docs	17
12	Statutory survey authority	39	31	Engine air pollution prevention	17
13	Dangerous goods compliance	38	32	Cargo Ship Safety Cert.	17
14	Anti-fouling systems	37	33	Minimum safe manning document	16
15	Continuous Synopsis Record	36	34	Navigation safety	16
16	Ballast water management	36	35	Propulsion & aux. machinery	16
17	Cargo Ship Safety Radio	29	36	Wreck removal convention (Nairobi)	16
18	Cargo Ship Safety Construction	29	37	IOPP	16
19	Detention duration	29	38	Fire safety	15

following analysis examines the predictive importance of individual PSC-derived features in greater detail. According to the overall feature ranking shown in Fig. 5 and Table 7, in the Ship Particulars category, Ship Type Value ranks second with a score of 70, indicating relatively high predictive importance in the fitted model. Gross Tonnage Value and Keel Date Value rank third and seventh with feature importance scores of 65 and 52, respectively, validating the continued significance of vessel size parameters and age as traditional risk indicators in data-driven models. In the PSC Inspection Information category, PSC Inspection Type ranks fourth with a feature importance score of 62, indicating that the gradient of regulatory inspections is a key predictive variable in risk assessment. Number of deficiencies and Flag Value rank eighth and ninth with weights of 48 and 47, respectively, confirming the important contribution of deficiency accumulation effects and the quality of flag state regulation in predicting accident severity. Class Certificates Issuing Authority and Statutory Certificates Surveying Authority Type received feature importance scores of 40 and 39, respectively. The former reflects the performance category associated with the relevant classification or recognized organization, while the latter distinguishes whether statutory survey and certification were undertaken

by the flag State Administration or by an authorized RO. Their relatively high importance may therefore reflect the regulatory and institutional information embedded in ship survey and certification arrangements. Among the Certification Documents features, the Maritime Labour Certificate (MLC) ranks first with an importance score of 43. The MLC is an international certificate for crew work and living conditions, covering areas such as crew contracts, wages, living environment, and medical care, and is an important institutional safeguard for ensuring crew human rights. Not all vessels are required to hold this certificate; it is typically mandatory for commercial vessels above a certain tonnage engaged in international voyages. The relatively high importance of the MLC may partly reflect the regulatory and operational profile represented by this certificate. Because its certification requirements apply to specified categories of commercially operated ships engaged in international operations, the presence of the MLC may also capture differences in vessel and trading profiles associated with its regulatory scope. Among the Deficiencies features, MLC stands out with a feature importance score of 24. MLC-related deficiencies reflect compliance information concerning seafarers' working and living conditions. Given the established safety relevance of human-element issues such as fatigue and work–rest arrangements, this provides a plausible explanation for why MLC-related deficiency information emerges as an informative predictor in the present model. The prominence of MLC-related information across both certification documents and deficiency categories further indicates its value in characterizing vessel safety conditions. Additionally, several certification document variables (e.g., Document of Compliance Dangerous Goods and International Anti-Fouling System), and deficiency-category indicators (e.g., Safety of Navigation, Propulsion and auxiliary machinery) yield importance scores between 16 and 38 points, indicating that these micro-indicators provide invaluable supplementary value in identifying potential systemic fault chains.

4.2.2. Comparative validation of selected feature contributions

For comparative validation of the feature patterns identified by the forward feature-selection procedure, SHAP analysis was additionally conducted using the LGBM model corresponding to the best-performing feature subset identified above. Using the same predictive model provides a consistent basis for comparing the features emphasized by the two analyses. Unlike split-count importance, which reflects how frequently a feature is used for tree splitting, SHAP values quantify the magnitude and direction of individual feature contributions to the model output [55]. The SHAP results are therefore used as complementary evidence to examine the consistency of the feature patterns identified through the proposed feature-selection procedure.

As shown in Fig. 6 and Table 7, the SHAP results are broadly consistent with the tree-based ranking for the principal predictive features. Four of the five highest-ranked features in the tree-based analysis—Location, Ship Type Value, Gross Tonnage Value, and Number of investigation reports—also appear among the five leading features according to mean absolute SHAP values. PSC Inspection Type, Flag Value, Keel Date Value, and Number of deficiencies also remain among the relatively important SHAP features. Overall, the two rankings show a clear positive correspondence, while some differences in the exact ordering are observed because split-count importance measures feature usage during tree construction, whereas SHAP quantifies contribution to model outputs. This complementary analysis provides additional support for the predictive relevance of the features identified by the proposed feature-selection procedure.

4.3. Theoretical interpretation and practical value

The feature importance analysis provides empirical evidence of predictive associations between PSC inspection data and maritime accident severity. The following discussion interprets these associations by considering the regulatory meaning of the identified variables and

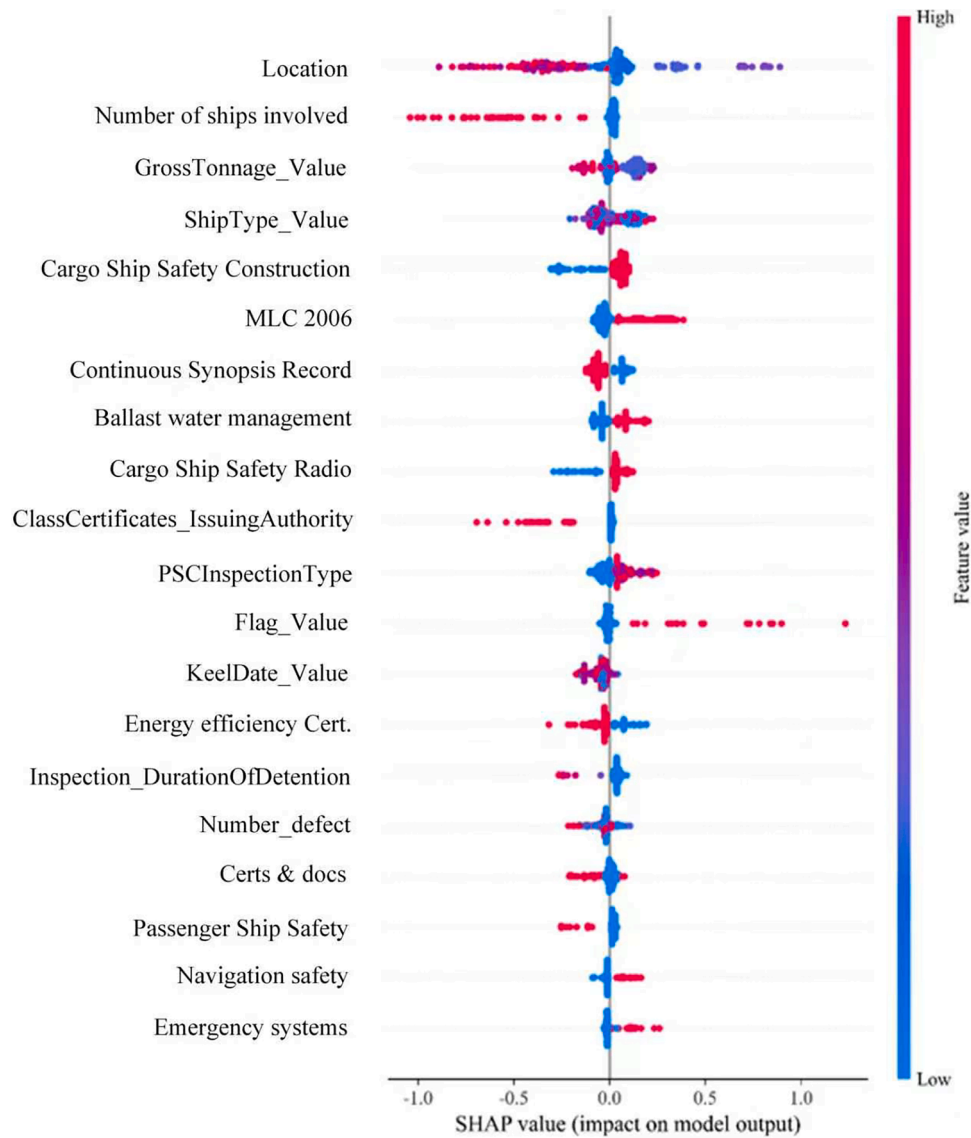


Fig. 6. Feature importance ranking based on SHAP analysis.

relevant maritime safety evidence, thereby providing a data-driven basis for risk-oriented interpretation and management implications.

The consistently high importance of Ship Type Value and Gross Tonnage Value in both the forward feature-selection ranking and SHAP analysis aligns with the theoretical framework of risk heterogeneity across vessel types [56]. Tankers and bulk carriers face a high risk of system failure due to operational intensity and specialized cargo; large vessels, due to operational inertia and system complexity, often experience more severe failure effects during accidents [1]. Similarly, the prominence of Keel Date is consistent with established safety concerns regarding vessel aging. These findings may provide supplementary evidence for risk-oriented PSC inspection prioritization, suggesting that vessel characteristics such as size and age could be considered together with existing regulatory criteria when interpreting model-derived risk information. The prominent performance of PSC inspection-related features reflects the value of regulatory information in accident-severity assessment. The high importance of PSC Inspection Type is consistent with the risk-based design of the Paris MoU inspection regime, in which inspection scope and priority are related to the Ship Risk Profile and to conditions identified during inspection. Inspection type may therefore embed part of the pre-existing regulatory risk assessment. Similarly, Number of deficiencies and Flag Value

correspond to historical and regulatory information incorporated into the Paris MoU risk framework. Their prominence in the present model suggests that these regulatory indicators retain useful information when the analytical focus is extended from PSC targeting to accident-severity assessment.

Most certification document and deficiency-category features fall within the medium importance range, indicating the potential supplementary value of these more specific risk indicators. MLC deficiency also exhibits a relatively strong contribution in the SHAP analysis, providing complementary evidence of its predictive relevance beyond its split-count ranking, consistent with the existed empirical research et al. [57], which suggests that fatigue and stress may relate to crew working conditions. This suggests that the relationship between crew working conditions and operational errors may warrant closer attention in future validation as a supplementary predictive signal. The importance of Safety of Navigation and Propulsion and auxiliary machinery supports the conclusions of another research [58], which state that the reliability of navigation systems and propulsion equipment integrity are critical for preventing severe accidents. These findings may inform future evaluation of inspection-priority and resource-allocation strategies.

In summary, the feature-importance results demonstrate that PSC inspection records provide supplementary predictive information for

accident severity assessment and offer a practical basis for risk-oriented decision support. The resulting model-derived evidence should be interpreted as supplementary risk information rather than as a stand-alone regulatory warning; it can help PSC authorities prioritize attention toward vessels and safety dimensions warranting closer consideration when determining ship-selection and inspection priorities. At the company level, the same evidence may support shipowners and management companies in prioritizing preventive actions and allocating safety-management resources.

4.4. Robustness, ablation, and complementary performance evaluation

To further evaluate the reliability and predictive performance of the proposed framework, this section conducts three complementary analyses. First, five-fold cross-validation is used to examine the robustness of the probability-error results against variations in data partitioning. Second, an ablation analysis evaluates the individual and combined contributions of PSC-accident data fusion, SMOTEN, CTDDPM, and feature selection to the overall performance improvement. Finally, conventional classification and discrimination metrics are reported for the final full framework as supplementary evidence of predictive performance. These additional classification metrics complement, rather than replace, the probability-error metrics adopted as the primary basis for model evaluation in this study.

4.4.1. Five-fold cross-validation of PSC information integration

To evaluate the stability of model performance across different data partitions, stratified five-fold cross-validation was adopted for all nine prediction models. The IMO accident-data-only setting and the PSC-accident fusion setting were evaluated using the same fold partitions, and the resulting MSE values are summarized in Table 8.

As shown in Table 8, the average MSE across the nine models decreased from 0.246596 ± 0.014353 for the IMO accident data alone to 0.239445 ± 0.008128 after incorporating PSC records, corresponding to an average reduction of approximately 2.90%. In addition, the smaller standard deviation observed after PSC integration indicates lower variation in the averaged prediction error across the five data partitions. The paired fold-level mean MSE reduction was 0.007151 (95% CI: -0.001872 to 0.016174). Although the interval included zero, the positive point estimate, together with the lower mean MSE observed for eight of the nine models in the initial five-fold evaluation, indicated a predominantly positive direction of improvement. To further assess the statistical significance of PSC information integration, the evaluation was extended to five random seeds with matched stratified five-fold partitions. Paired *t*-tests were performed using five pairs of seed-wise mean MSE values, each averaged across the five folds and nine prediction models. As reported in Table C5 of Appendix C, PSC integration reduced the average MSE from 0.245040 ± 0.001350 to 0.239619 ± 0.000838 , with a statistically significant improvement confirmed by a two-sided paired *t*-test ($p = 0.0009$). Together, these results provide further statistical evidence for the supplementary

predictive value of PSC inspection records in maritime accident severity prediction.

4.4.2. Ablation analysis of the proposed framework

To systematically evaluate the contributions and synergies of each component in the multi-stage data augmentation and feature selection framework, a series of ablation experiments were conducted, with MSE results shown in Fig. 7. Supplementary evaluation metrics, such as RMSE, MAE and Log Loss, are provided in Tables C1–C3 in Appendix C for reference. Building on the PSC-fusion setting evaluated in Section 4.4.1, the ablation analysis further examines the contributions of the data-augmentation and feature-selection components to the proposed framework. Fig. 7 presents the MSE results obtained under different combinations of PSC information integration, data augmentation, and feature selection. For LGBM, incorporating PSC inspection records reduced the MSE from 0.243048 to 0.241150, corresponding to a reduction of approximately 0.78%. Although this improvement was modest, it provides empirical evidence that PSC inspection records contribute supplementary predictive information beyond the conventional accident variables. The comparative experiments in Section 4.1 further showed that CTDDPM achieved the lowest average downstream MSE among CTGAN, TVAE, and TabDDPM under the common 2000-sample augmentation setting, supporting its use as the refined generative stage of the proposed augmentation framework. To further evaluate the benefit of the proposed two-stage augmentation strategy against a simpler sampling approach, random oversampling (ROS) was additionally introduced as a replication-based comparator. ROS was applied only to the training set by randomly replicating minority-class samples until class balance was reached, while the test set remained unchanged and the same fixed configurations of the nine prediction models were retained.

As a simple sampling comparator, ROS produced mixed model-specific effects, with an average MSE of 0.2550 across the nine prediction models. In comparison, the proposed two-stage SMOTEN+CTDDPM augmentation achieved a lower average MSE of 0.2347. The corresponding model-specific MSE results are reported in Table C4 of Appendix C. This result provides additional evidence that the predictive improvement of the proposed augmentation strategy is not attributable merely to simple replication of minority-class training samples. As shown in Fig. 8, the independent effectiveness of SMOTEN, CTDDPM, and FS is further examined using the PSC-accident fused data. SMOTEN, when applied alone, alleviates the small-sample learning dilemma by compensating for class imbalance and expanding the training sample boundary, reducing the average MSE by approximately 1.92% compared to the original IMO accident database. When applied alone, CTDDPM's advanced generative mechanism learns the underlying data distribution and generates high-quality synthetic samples, increasing data volume and capturing complex feature dependencies, which reduces the average MSE by approximately 3.11%. Similarly, executing FS alone eliminates redundant and noisy features, optimizing the model's input representation and reducing the average MSE by approximately 4.05%. These results clearly show that each technique in the framework contributes positively and independently to improving model performance. This study also examines the synergistic effects among these strategies. When SMOTEN, CTDDPM, and FS are combined in pairs, each pairwise configuration achieves a lower average MSE than either of its corresponding single-component settings, indicating a complementary effect among the three components. Compared to the original IMO accident database, the average MSE decreases by 5.16% and the average Log Loss decreases by 5.96%, indicating a positive cumulative enhancement effect, rather than negative interaction, between these strategies. When the three methods, SMOTEN, CTDDPM, and FS, are applied together, the full framework achieves the lowest average MSE among all evaluated settings. Notably, the full framework reduced the LGBM MSE from 0.243048 to 0.215787, corresponding to an approximately 11.22% reduction relative to the baseline and a further 10.52% reduction

Table 8
Five-fold cross-validation results before and after PSC information integration.

Model	IMO accident data only MSE (mean $\pm$ SD)	IMO + PSC records MSE (mean $\pm$ SD)
LGBM	0.225955 $\pm$ 0.022690	0.223429 $\pm$ 0.011270
XGBoost	0.230260 $\pm$ 0.021340	0.225990 $\pm$ 0.011608
GBDT	0.228765 $\pm$ 0.019116	0.224841 $\pm$ 0.011807
CatBoost	0.227344 $\pm$ 0.019032	0.227782 $\pm$ 0.011587
AdaBoost	0.252540 $\pm$ 0.009370	0.250536 $\pm$ 0.013453
Random Forest	0.230289 $\pm$ 0.012137	0.230201 $\pm$ 0.010226
TabNet	0.350163 $\pm$ 0.018507	0.300471 $\pm$ 0.018643
Neural Network	0.237666 $\pm$ 0.015473	0.235556 $\pm$ 0.009084
SVM	0.236376 $\pm$ 0.016494	0.236197 $\pm$ 0.006333
Average across nine models	0.246596 $\pm$ 0.014353	0.239445 $\pm$ 0.008128

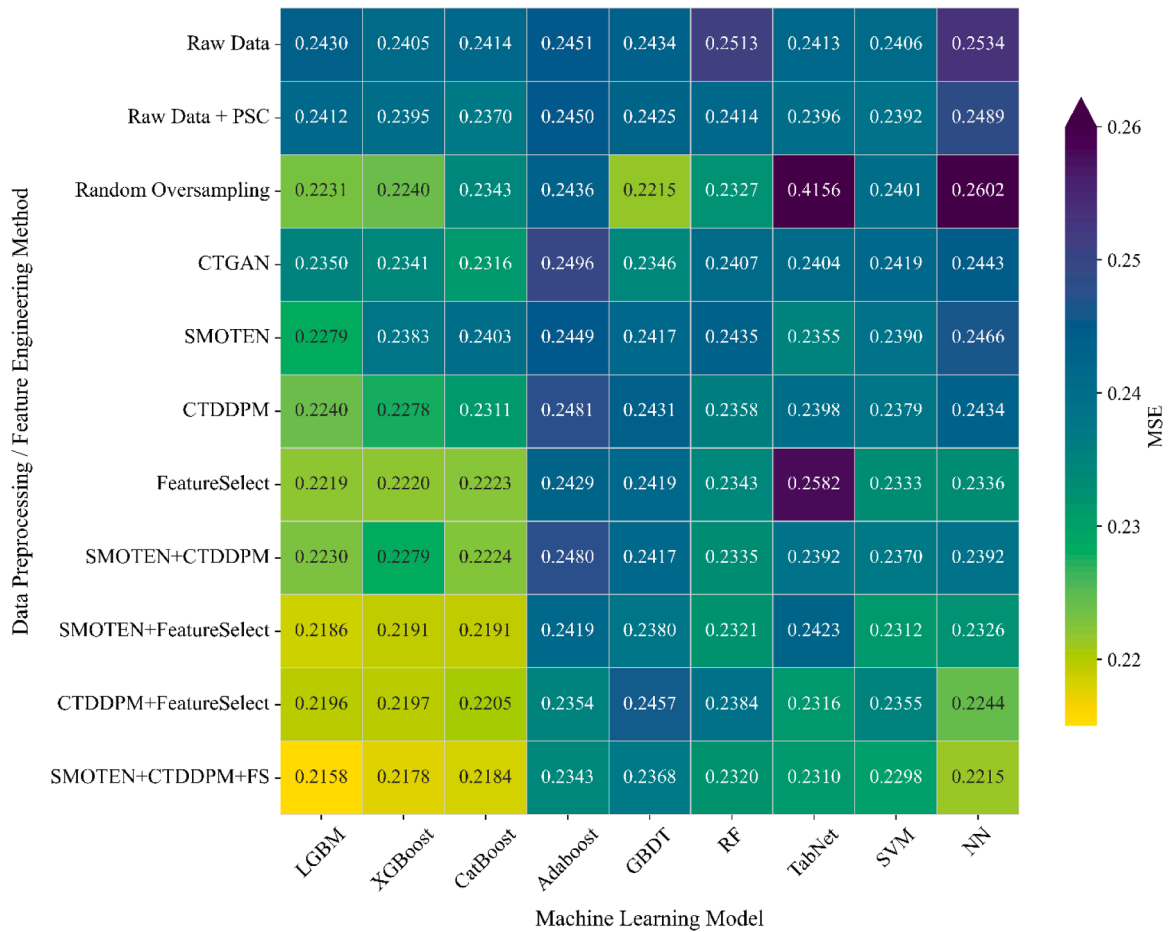


Fig. 7. MSE-based ablation study of data augmentation components.

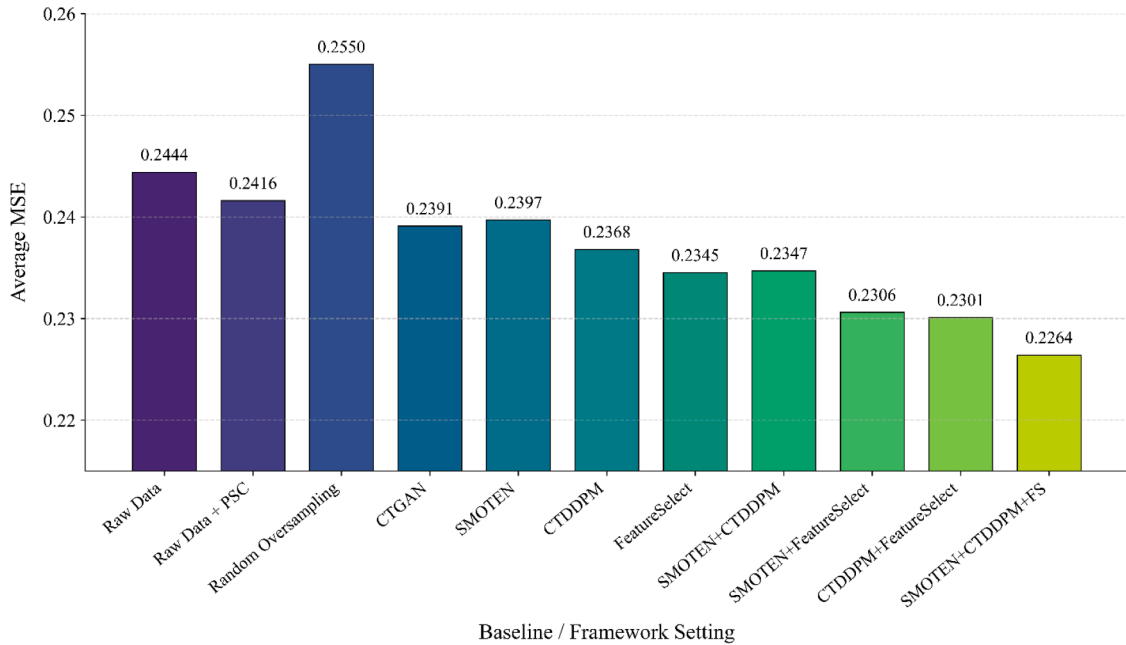


Fig. 8. Average MSE across different baseline and framework settings.

relative to the PSC–accident fusion-only setting, which did not include SMOTEN, CTDDPM, or feature selection. This result further supports the overall effectiveness of the proposed multi-stage data augmentation and

feature-selection framework. This improvement was further confirmed by the five-seed cross-validation results presented in Table C5 of Appendix C. The complete framework achieved an average MSE of

0.234840 ± 0.001123 , compared with 0.239619 ± 0.000838 for PSC fusion alone, corresponding to a further reduction of approximately 1.99% (two-sided paired *t*-test, $p = 0.0010$).

The examination of different model architectures under the optimal augmentation strategy provides further insights. Advanced ensemble models based on Gradient Boosting Decision Trees (GBDT), such as LGBM, XGBoost, and CatBoost, demonstrate outstanding performance due to their powerful handling of tabular data and intrinsic regularization mechanisms, consistent with the existed findings et al. [19] regarding LGBM's superiority in predicting maritime accident severity. In contrast, AdaBoost, RF, and SVM models perform relatively poorly, possibly due to their sensitivity to noise, difficulty in capturing complex interactions between features, and the generalization capability of kernel functions in high-dimensional categorical data [24]. Notably, the Neural Network model, despite having a higher error on the original data (MSE = 0.2534), reduced to 0.2215 after data augmentation and feature optimization, a 12.56% reduction, further validating the key role of high-quality data generation in enhancing deep learning model performance under small sample conditions.

The results of the ablation study systematically validate the foundational role of PSC-accident data fusion, as well as the independent value and synergistic effects of SMOTEN, CTDDPM, and FS in improving maritime accident-severity prediction model performance. The multi-stage strategy addresses the challenges of small sample sizes and high-dimensional categorical data. CTDDPM, as the core enhancement method, generates synthetic data crucial for the model's learning of robust decision boundaries, both in quantity and quality. It highlights the need to balance generation scale and model performance in practical applications. Although data augmentation improves the performance of all models, GBDT models like LGBM remain the preferred choice due to their adaptability to the current data characteristics. The data augmentation framework proposed in this study, particularly the use of conditional diffusion models, alleviates the model's generalization bottleneck under small sample conditions and provides an empirically supported augmentation path for high-risk, low-sample applications like accident-severity prediction.

4.4.3. Complementary classification performance evaluation

The preceding robustness and ablation analyses primarily evaluate model performance using probability-error metrics, as the present study focuses on the quality of probabilistic accident-severity estimates rather than solely on threshold-based class assignments. To provide complementary evaluation using conventional classification measures, Accuracy, macro-F1, ROC-AUC, PR-AUC, sensitivity, specificity, and balanced accuracy were additionally calculated for the nine prediction models under the final SMOTEN+CTDDPM+FS framework.

As shown in Table 9, the supplementary classification metrics provide additional evidence for the predictive performance of the final framework while preserving the probability-error-based evaluation as the primary comparative basis of this study. Macro-F1 is reported to give equal consideration to the two severity classes, while sensitivity specifically represents the recall of Class 1, corresponding to severe accidents. ROC-AUC and PR-AUC are calculated from the continuous

prediction scores and therefore provide complementary information on model discrimination beyond threshold-based classification results.

LGBM and XGBoost both achieved the highest Accuracy of 0.7200, with macro-F1 values of 0.7200 and 0.7194, respectively. XGBoost achieved the highest ROC-AUC of 0.7199, while LGBM maintained strong and comparatively balanced classification performance, with a sensitivity of 0.7273 and a specificity of 0.7130. These supplementary results are consistent with the primary probability-error-based evaluation and further support LGBM as the best-performing and representative model under the evaluation framework adopted in this study. Accordingly, the classification behaviour and probability calibration of LGBM were further examined using a confusion matrix and a calibration curve constructed with five equal-frequency bins, as shown in Figs. E1 and E2 in Appendix E, respectively. The confusion matrix further confirms the balanced classification capability of LGBM across the two severity classes, while the calibration curve provides complementary evidence regarding the consistency between predicted probabilities and observed outcomes. Together, these diagnostic results provide additional support for the reliability of LGBM under the final framework.

5. Conclusions

Accurate prediction of maritime accident severity is crucial for proactive safety measures and risk management. While PSC inspection records have primarily been used for regulatory purposes, their potential in predicting accident severity has not been fully explored. This study bridges this gap by integrating PSC inspection record data with the IMO accident database to create a dynamic, forward-looking dataset that combines pre-incident regulatory information and post-incident data. This data fusion enables a shift from reactive to proactive maritime safety management. This study proposes a PSC-accident multi-source fusion framework using data augmentation to address challenges such as high-dimensionality and data scarcity. SMOTEN alleviates sample scarcity, CTDDPM generates high-quality synthetic data with stronger label consistency, and feature selection enhances model performance and interpretability. Empirical results show that integrating PSC inspection records reduced the average MSE across the nine models by 1.15%; under the full framework, LGBM achieved an 11.22% MSE reduction relative to its accident-data-only baseline.

Feature importance analysis identifies vessel type, as recorded in PSC inspection records, as a key predictive factor associated with accident severity. Notably, both the Maritime Labour Certificate and MLC-related deficiencies emerge as important predictive features. The former represents the MLC certification and compliance dimension concerning seafarers' working and living conditions, while the latter reflects deficiencies identified in areas such as conditions of employment, work and rest arrangements, accommodation, and health protection. Their joint prominence suggests that MLC-related certification and deficiency information provides complementary predictive information for accident-severity assessment. CTDDPM also achieved the lowest average MSE among the compared generative methods, suggesting that the moderate perturbations introduced by its diffusion process may contribute to improved downstream prediction. The PSC-accident multi-

Table 9

Supplementary classification performance under the final SMOTEN+CTDDPM+FS framework.

Model	Accuracy	Macro-F1	ROC-AUC	PR-AUC	Sensitivity	Specificity	Balanced Accuracy
LGBM	0.7200	0.7200	0.7076	0.6586	0.7273	0.7130	0.7202
XGBoost	0.7200	0.7194	0.7199	0.6273	0.7818	0.6609	0.7213
AdaBoost	0.6667	0.6666	0.6585	0.6169	0.7000	0.6348	0.6674
GBDT	0.7111	0.7110	0.6977	0.6397	0.7455	0.6783	0.7119
CatBoost	0.6089	0.6083	0.5735	0.5558	0.6636	0.5565	0.6101
RF	0.6800	0.6800	0.6964	0.6792	0.7000	0.6609	0.6804
TabNet	0.6444	0.6444	0.6532	0.5895	0.6727	0.6174	0.6451
SVM	0.6711	0.6620	0.7085	0.6336	0.8545	0.4957	0.6751
Neural Network	0.6533	0.6525	0.6667	0.6079	0.7182	0.5913	0.6547

source fusion framework enhances model performance and may provide preliminary evidence for future policy evaluation by regulatory agencies. The findings may inform future evaluation of the weight configuration of variables in the ship risk profile mechanism, potentially supporting the identification of ageing vessels and vessels with frequent structural deficiencies. The certificate and deficiency patterns identified by the model may also provide supplementary evidence for considering future inspection priorities.

While this study presents a novel approach to maritime accident severity prediction by integrating PSC inspection records with accident data, there are several limitations that warrant consideration. On one hand, this research primarily uses PSC inspection records from the Paris MoU. Although regional PSC regimes operate under broadly harmonized international regulatory principles, the empirical transferability of the present framework to vessels represented exclusively in other regional MoUs has not yet been directly validated. Moreover, because inclusion in the fused dataset requires an eligible Paris MoU inspection record, the matched accident sample represents a selected subset of the global GISIS accident records rather than the full population of accident-involved vessels. Consequently, vessels without recent Paris MoU inspection histories, particularly those operating primarily under other regional PSC regimes, may be less represented in the matched sample, which may affect the observed feature distributions and the applicability of the findings beyond the Paris MoU context. To enhance the applicability of the framework, future research could incorporate PSC inspection records from multiple regional MoUs (e.g., Tokyo MoU, Indian Ocean MoU). On the other hand, the performance of the proposed PSC-accident multi-source fusion framework is currently limited by the size of the integrated dataset. Although the current framework demonstrates promising results, the dataset is relatively small size which may constrain the model's potential, particularly in capturing the full complexity of maritime accident severity. Expanding the dataset by

incorporating more extensive and diverse accident records, as well as additional PSC inspection data over an extended period, could enhance the framework's performance.

CRediT authorship contribution statement

Tianyi Li: Writing – original draft, Visualization, Methodology, Data curation. **Xinjian Wang:** Writing – original draft, Investigation, Formal analysis, Conceptualization. **Chia-Hsun Chang:** Writing – original draft, Resources, Investigation, Formal analysis. **Manel Grifoll:** Writing – review & editing, Validation, Investigation, Formal analysis. **Hongxiang Feng:** Validation, Resources, Investigation, Formal analysis. **Zaili Yang:** Writing – original draft, Validation, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known financial conflicts of interest or personal relationships that could have influenced the work presented in this paper.

Acknowledgements

The authors gratefully acknowledge support from the National Natural Science Foundation of China [Grant No 52671450], the Fundamental Research Funds for the Central Universities [Grant No 3132026122; 3132026522], the Joint Project of Liaoning Provincial Science and Technology Program [Grant No 2025-BSLH-058] and the International Exchanges program of Royal Society, UK [Grant No IEC \NSFC\233298] and the European Research Council project under the European Union's Horizon 2020 research and innovation programme [Grant No TRUST CoG 2019 864724].

Appendix A. Feature Documentation of the Accident Dataset

Table A1
Field Descriptions of the IMO Maritime Casualty Dataset.

Column Name	Description	Column Name	Description
Number Of Ships Involved	Number of ships involved in the incident.	Classification Society 1	Classification society responsible for the first ship.
Ships Involved	Names or identifiers of the ships involved.	Dead weight 1	Deadweight tonnage (DWT) of the first ship.
SOLAS Status	Whether the ship falls under the scope of the SOLAS Convention.	Cargo 1	Type or description of cargo carried by the first ship.
Flag States	Flag States (countries of registration) of the involved ships.	Crew On board 1	Number of crew members onboard the first ship at the time of the incident.
Incident Date Time	Date and time when the incident occurred.	Passengers On board 1	Number of passengers onboard the first ship.
Type Of Casualty	Type of casualty (e.g., collision, grounding, fire).	Others On board 1	Number of other personnel onboard the first ship (e.g., inspectors, third parties).
Coordinates	Geographical coordinates (latitude and longitude) of the incident.	Dead Missing Crew 2	Number of crew members reported dead or missing.
Location	Descriptive geographical or administrative location of the incident.	Dead Missing Passengers 2	Number of passengers reported dead or missing.
Number Of Investigation Reports	Number of investigation reports published for this incident.	Dead Missing Others 3	Number of other personnel reported dead or missing.
Number Of Analyses	Number of analytical reports associated with the incident.	Seriously Injured Crew 2	Number of crew members seriously injured.
Initial Event	Initial triggering event or cause of the incident.	Seriously Injured Passengers 2	Number of passengers seriously injured.
Summary Of Events	Narrative summary of the incident sequence.	Seriously Injured Others 2	Number of other personnel seriously injured.
ShipType 1	Type of the first ship involved (e.g., bulk carrier, tanker, container ship).	Investation Report Date	Date when the official investigation report was published.
Consequences 1	Consequences suffered by the first ship (e.g., sinking, fire, damage).	Analysis Date	Date when any analytical report was published.
Gross Tonnage 1	Gross tonnage (GT) of the first ship, indicating overall size.		

Table A2
High-Level Categories of PSC Deficiency Codes.

Deficiency Code	Description	Deficiency Code	Description
01	Certificates & Documentation	10	Safety of Navigation
02	Structural condition	11	Life-saving appliances
03	Water/Weather tight condition	12	Dangerous Goods
04	Emergency Systems	13	Propulsion and auxiliary machinery
05	Radio communication	14	Pollution Prevention
06	Cargo operations including equipment	15	ISM
07	Fire safety	16	ISPS
08	Alarms	18	MLC, 2006
09	Working and Living Conditions	99	Other

Table A3
Detailed Categories of PSC Certificate Codes.

Code	Description	Code	Description
501	Cargo Ship Safety Construction	519	High Speed Craft Safety
502	Cargo Ship Safety Equipment	520	INF Certificate of Fitness
503	Cargo Ship Safety Radio	521	Mobile Offshore Drilling Unit Safety
504	Cargo Ship Safety	522	Special Purpose Ship Safety
505	International Oil Pollution Prevention (IOPP)	523	Tonnage
506	International Air Pollution Prevention	524	Other Certificate
507	International Sewage Pollution Prevention	525	Continuous Synopsis Record
508	Load Line	526	Engine International Air Pollution Prevention
509	Document of Compliance	527	Document of Compliance Dangerous Goods
510	Safety Management Certificate	528	International Ballast Water Management
511	International Ship Security	529	International Anti-Fouling System
512	Minimum Safe Manning Document	530	Civil Liability for Oil Pollution Damage
513	Passenger Ship Safety	531	Bunker Oil Pollution Damage
514	Dangerous Chemicals in Bulk (CoF/BC Code)	532	International Energy Efficiency Certificate
515	Dangerous Chemicals in Bulk (ICoF/IBC Code)	533	Maritime Labour Certificate
516	Liquified Gases in Bulk (CoF/GC Code)	535	Polar Ship Certificate
517	Liquified Gases in Bulk (ICoF/IGC Code)	536	Nairobi International Convention on the Removal of Wrecks
518	International Pollution Prevention (NLS in Bulk)		

Appendix B. Algorithmic Workflow of CTDDPM

Algorithm B1
ConditionalTabDDPM Model.

```

Input: X_train; y_train; num_timesteps; epochs; batch_size
Output: generated_samples
1      Begin
2      Initialize model components:
3      label_embed: Label embedding layer
4      time_embed: Time step embedding layer
5      model: Main generation network
6      noise_scheduler: Noise scheduler (DDPMScheduler)
7      discrete_values: Predefined set of discrete values
8      For each epoch in 1 to epochs:
9          Set total_loss = 0
10         For each batch_X, batch_y in X_train, y_train:
11             Select random time steps (timesteps) for the batch
12             For each time step t in timesteps:
13                 Add noise to batch_X based on t using noise_scheduler
14             Predict the noise (pred_noise) using the model:
15             - Pass noisy_X, t, and batch_y through the model
16             Compute MSE loss (mse_loss) between pred_noise and actual noise
17             Compute discrete loss (disc_loss) based on pred_noise and discrete_values
18             Calculate total_loss = mse_loss + lambda_disc * disc_loss
19             Perform backpropagation and update model parameters
20             Add current loss to total_loss
21         For each generated sample i in 1 to num_samples:
22             Initialize sample with random noise
23             For each time step t in reversed(range(0, num_timesteps)):
24                 For each label y_cond in batch_y:
25                     Predict noise (pred_noise) for sample, t, and y_cond using the model
26                     Use noise_scheduler to remove noise from sample, obtaining prev_sample

```

(continued on next page)

Algorithm B1 (continued)

27	Map final prev_sample to discrete values using discrete_values
28	Add generated sample to generated_samples
29	Return generated_samples
30	End

Appendix C. Evaluation Metrics under Data Augmentation Settings**Table C1**

RMSE under Different Augmentation Strategies.

RMSE	LGBM	XGBoost	Adaboost	GBDT	CatBoost	RF	TabNet	SVM	neural_network
Raw accident data	0.4930	0.4904	0.4951	0.4933	0.4913	0.5013	0.4912	0.4905	0.5034
Raw accident data+PSC data	0.4911	0.4894	0.4950	0.4924	0.4869	0.4913	0.4895	0.4891	0.4989
CTGAN	0.4847	0.4838	0.4996	0.4843	0.4812	0.4906	0.4903	0.4919	0.4942
SMOTEN	0.4774	0.4882	0.4949	0.4917	0.4902	0.4935	0.4853	0.4889	0.4966
CTDDPM	0.4733	0.4773	0.4981	0.4930	0.4807	0.4856	0.5016	0.4878	0.4897
FeatureSelect	0.4711	0.4711	0.4928	0.4919	0.4715	0.4840	0.5081	0.4831	0.4833
SMOTEN+CTDDPM	0.4722	0.4774	0.4979	0.4916	0.4716	0.4832	0.4929	0.4869	0.4891
SMOTEN+FeatureSelect	0.4686	0.4681	0.4919	0.4878	0.4681	0.4818	0.4922	0.4809	0.4823
CTDDPM+FeatureSelect	0.4676	0.4687	0.4852	0.4957	0.4696	0.4883	0.4812	0.4853	0.4737
SMOTEN+CTDDPM+FS	0.4645	0.4667	0.4840	0.4866	0.4673	0.4816	0.4807	0.4794	0.4706

Table C2

MAE under Different Augmentation Strategies.

MAE	LGBM	XGBoost	Adaboost	GBDT	CatBoost	RF	TabNet	SVM	neural_network
Raw accident data	0.4894	0.4884	0.4919	0.4861	0.4896	0.4437	0.4579	0.4784	0.4509
Raw accident data+PSC data	0.4301	0.4345	0.4946	0.4240	0.4449	0.4566	0.4806	0.4795	0.4637
CTGAN	0.4569	0.4623	0.4996	0.4752	0.4727	0.4696	0.4712	0.4898	0.4899
SMOTEN	0.4365	0.4392	0.4946	0.4427	0.4505	0.4597	0.4782	0.4772	0.4845
CTDDPM	0.4385	0.4485	0.4981	0.4425	0.4548	0.4792	0.4997	0.4839	0.4537
FeatureSelect	0.4428	0.4414	0.4900	0.4751	0.4435	0.4746	0.4898	0.4651	0.4609
SMOTEN+CTDDPM	0.4552	0.4475	0.4980	0.4432	0.4470	0.4794	0.4854	0.4818	0.4449
SMOTEN+FeatureSelect	0.4361	0.4395	0.4891	0.4737	0.4353	0.4746	0.4744	0.4712	0.4599
CTDDPM+FeatureSelect	0.4417	0.4360	0.4743	0.4942	0.4481	0.4847	0.4671	0.4825	0.4591
SMOTEN+CTDDPM+FS	0.4390	0.4334	0.4738	0.4778	0.4497	0.4753	0.4569	0.4732	0.4378

Table C3

Logarithmic Loss under Different Augmentation Strategies.

LogLoss	LGBM	XGBoost	Adaboost	GBDT	CatBoost	RF	TabNet	SVM	neural_network
Raw accident data	0.6791	0.6740	0.6832	0.6799	0.6759	0.7774	0.6931	0.6741	0.7588
Raw accident data+PSC data	0.6777	0.6810	0.6831	0.7010	0.6764	0.6797	0.6711	0.6716	0.7003
CTGAN	0.6633	0.6613	0.6924	0.6614	0.6555	0.6767	0.6732	0.6769	0.6818
SMOTEN	0.6498	0.6777	0.6829	0.6875	0.6853	0.6892	0.6618	0.6713	0.6879
CTDDPM	0.6430	0.6519	0.6894	0.6973	0.6547	0.6644	0.6965	0.6689	0.6759
FeatureSelect	0.6353	0.6375	0.6789	0.6796	0.6380	0.6608	0.7532	0.6594	0.6607
SMOTEN+CTDDPM	0.6383	0.6516	0.6892	0.6922	0.6394	0.6596	0.6781	0.6670	0.6771
SMOTEN+FeatureSelect	0.6293	0.6298	0.6768	0.6707	0.6311	0.6563	0.6889	0.6550	0.6567
CTDDPM+FeatureSelect	0.6290	0.6347	0.6643	0.6846	0.6330	0.6698	0.6547	0.6640	0.6401
SMOTEN+CTDDPM+FS	0.6231	0.6320	0.6610	0.6683	0.6273	0.6565	0.6496	0.6523	0.6329

Table C4

Detailed MSE values under different baseline and framework settings.

Setting	LGBM	XGBoost	CatBoost	AdaBoost	GBDT	RF	TabNet	SVM	NN	Mean
Raw accident data	0.2430	0.2405	0.2414	0.2451	0.2434	0.2513	0.2413	0.2406	0.2534	0.2444
Raw accident data + PSC data	0.2412	0.2395	0.2370	0.2450	0.2425	0.2414	0.2396	0.2392	0.2489	0.2416
Random oversampling	0.2231	0.2240	0.2343	0.2436	0.2215	0.2327	0.4156	0.2401	0.2602	0.2550
CTGAN	0.2350	0.2341	0.2316	0.2496	0.2346	0.2407	0.2404	0.2419	0.2443	0.2391
SMOTEN	0.2279	0.2383	0.2403	0.2449	0.2417	0.2435	0.2355	0.2390	0.2466	0.2397
CTDDPM	0.2240	0.2278	0.2311	0.2481	0.2431	0.2358	0.2398	0.2379	0.2434	0.2368

(continued on next page)

Table C4 (continued)

Setting	LGBM	XGBoost	CatBoost	AdaBoost	GBDT	RF	TabNet	SVM	NN	Mean
Feature Select	0.2219	0.2220	0.2223	0.2429	0.2419	0.2343	0.2582	0.2333	0.2336	0.2345
SMOTEN + CTDDPM	0.2230	0.2279	0.2224	0.2480	0.2417	0.2335	0.2392	0.2370	0.2392	0.2347
SMOTEN + FeatureSelect	0.2186	0.2191	0.2191	0.2419	0.2380	0.2321	0.2423	0.2312	0.2326	0.2306
CTDDPM + FeatureSelect	0.2196	0.2197	0.2205	0.2354	0.2457	0.2384	0.2316	0.2355	0.2244	0.2301
SMOTEN + CTDDPM + FS	0.2158	0.2178	0.2184	0.2343	0.2368	0.2320	0.2310	0.2298	0.2215	0.2264

Table C5

Multi-seed cross-validation results and paired significance tests for different framework settings.

Framework setting	MSE (mean $\pm$ SD)	Relative MSE reduction (%)	Paired t-test p-value
IMO accident data only	0.245040 $\pm$ 0.001350	—	—
IMO + PSC records	0.239619 $\pm$ 0.000838	2.21	0.0009
Full framework (SMOTEN+CTDDPM+FS)	0.234840 $\pm$ 0.001123	1.99	0.0010

Appendix D. Hyperparameter Settings and Supplementary Generative-Model Evaluation

Table D1

Key Hyperparameter Settings of the Prediction Models Used in the Final Experiments.

Model	Key Hyperparameter Settings
LGBM	n_estimators = 62; learning_rate = 0.17; max_depth = 3
XGBoost	n_estimators = 51; learning_rate = 0.18; max_depth = 4; eval_metric = logloss
GBDT	n_estimators = 60; learning_rate = 0.20; max_depth = 2
CatBoost	iterations = 68; learning_rate = 0.19; depth = 3
AdaBoost	n_estimators = 67; learning_rate = 0.19; base_estimator max_depth = 4
RF	n_estimators = 64; max_depth = 3
TabNet	n_d = 32; n_a = 32; n_steps = 3; gamma = 1.2; max_epochs = 100; patience = 10; batch_size = 1024; virtual_batch_size = 128; eval_metric = logloss
Neural Network	hidden_dim = 32; learning_rate = 0.01; optimizer = Adam; loss = BCELoss; batch_size = 256; epochs = 50; hidden activation = ReLU; output activation = Sigmoid
SVM	kernel = RBF; C = 0.1; gamma = scale; probability = True; preprocessing = StandardScaler

Table D2

Synthetic-Data Fidelity of Different Generative Models at the 2000-Sample Baseline.

Generative Model	Column Shapes Score	Column Pair Trends Score	Overall Score
CTDDPM	82.78	93.29	88.04
CTGAN	94.21	93.97	94.09
TVAE	91.99	93.84	92.91
TabDDPM	86.02	91.46	88.74

Table D3

Downstream Predictive Performance of Different Generative Augmentation Methods at the 2000-Sample Baseline.

Generative Model	CTGAN	CTDDPM	TVAE	TabDDPM
LGBM	0.235	0.224	0.2454	0.2306
XGBoost	0.2341	0.2278	0.2503	0.2271
GBDT	0.2346	0.2431	0.2453	0.235
CatBoost	0.2316	0.2311	0.2502	0.2433
AdaBoost	0.2496	0.2481	0.2396	0.2532
RF	0.2407	0.2358	0.2271	0.2451
TabNet	0.2404	0.2398	0.2499	0.2407
Neural Network	0.2443	0.2434	0.274	0.2497
SVM	0.2419	0.2379	0.2625	0.2467
Average MSE	0.2391	0.2368	0.2494	0.2413

Table D4

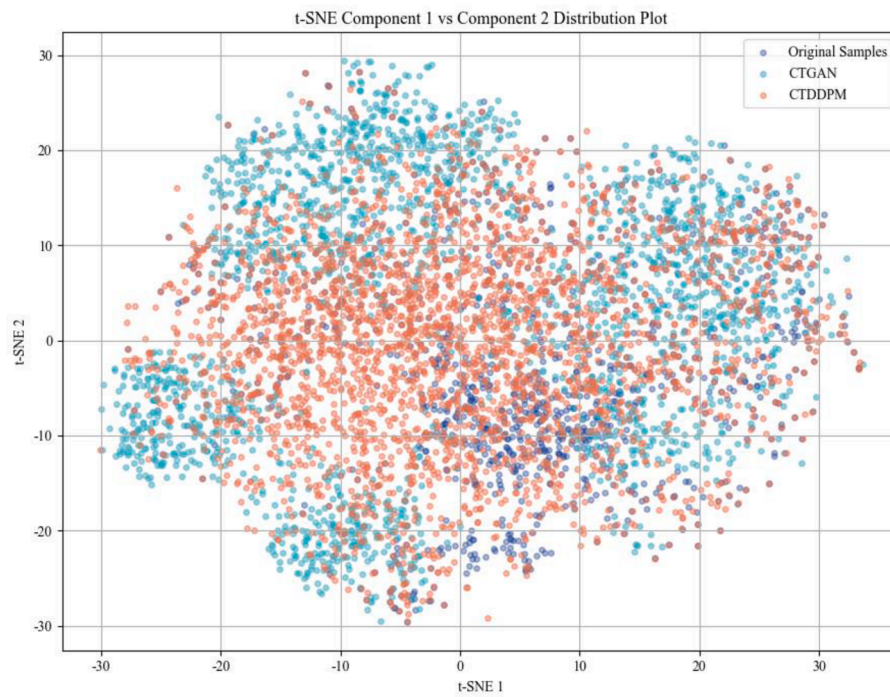
Multi-seed comparison and paired significance test of TabDDPM and CTDDPM.

Generative model	MSE (mean $\pm$ SD)	Relative MSE reduction (%)	Paired <i>t</i> -test p-value
TabDDPM	0.239360 $\pm$ 0.001596	—	—
CTDDPM	0.232471 $\pm$ 0.003817	2.88	0.0182

Table D5

Sensitivity analysis of key CTDDPM hyperparameters.

Hyperparameter	Value	MSE	Relative change (%)
Default configuration	1500 steps; $\lambda=1.0$; hidden=64	0.2389	—
Diffusion timesteps	500	0.2411	+0.93
	1000	0.2429	+1.68
Discrete regularization (λ)	0.5	0.2350	-1.62
	2.0	0.2459	+2.91
Hidden dimension	32	0.2435	+1.90
	128	0.2400	+0.47

**Fig. D1.** t-SNE Visualization of Real, CTGAN-Generated, and CTDDPM-Generated Samples.

Appendix E. Supplementary Diagnostic Evaluation of LightGBM

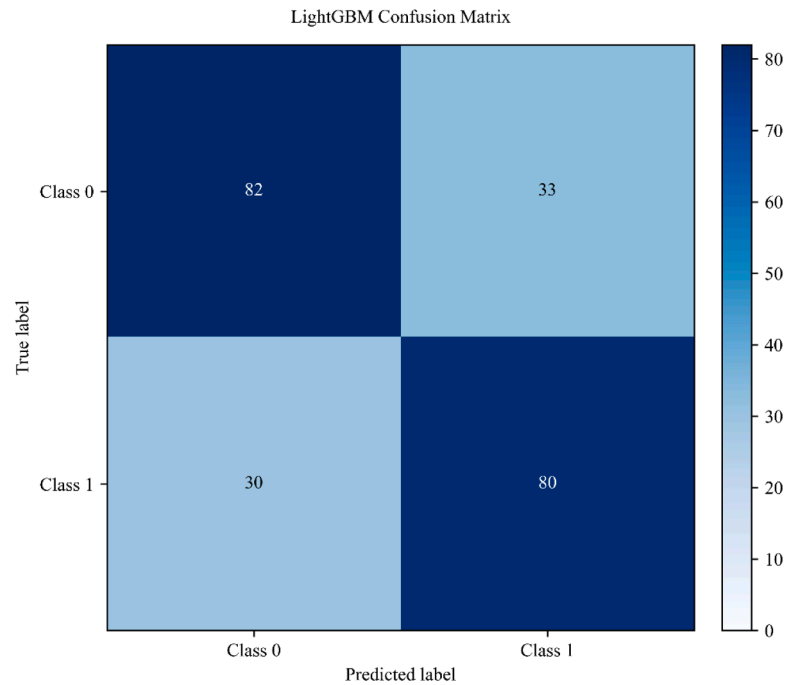


Fig. E1. Confusion matrix for LightGBM under the final framework on the holdout test set.

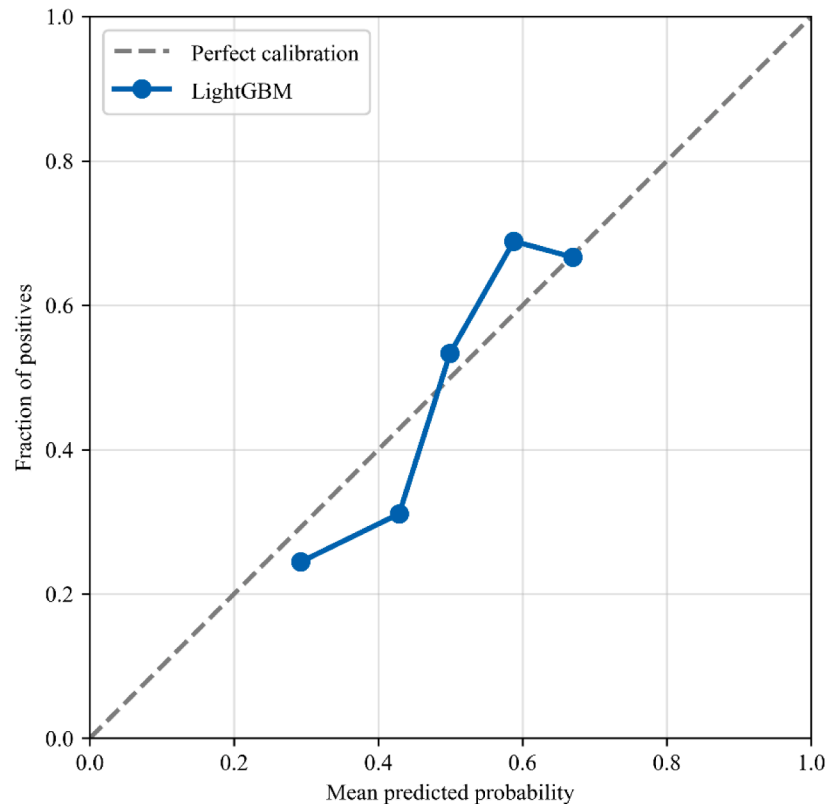


Fig. E2. LightGBM calibration curve with five equal-frequency bins.

Data availability

[com/AdvMarTech/Mari_Accident_PSC_Fusion](https://github.com/AdvMarTech/Mari_Accident_PSC_Fusion).

The data and source code is publicly available at: [https://github](https://github.com/AdvMarTech/Mari_Accident_PSC_Fusion).

References

- [1] Wang X, Cao W, Li T, Feng Y, Uğurlu Ö, Wang J. An integrated multidimensional model for heterogeneity analysis of maritime accidents during different watchkeeping periods. *Ocean Coast Manag* 2025;264:107625. <https://doi.org/10.1016/j.ocecoaman.2025.107625>.
- [2] Cao Y, Iulia M, Majumdar A, Feng Y, Xin X, Wang X, Wang H, Yang Z. Investigation of the risk influential factors of maritime accidents: a novel topology and robustness analytical framework. *Reliab Eng Syst Saf* 2025;254:110636. <https://doi.org/10.1016/j.res.2024.110636>.
- [3] European maritime safety agency. Annual overview of marine casualties and incidents 2024. EMSA; 2025.
- [4] Brandt P, Munim ZH, Chaal M, Kang H-S. Maritime accident risk prediction integrating weather data using machine learning. *Transp Res D: Transp Environ* 2024;136:104388. <https://doi.org/10.1016/j.trd.2024.104388>.
- [5] Cao W, Wang X, Zhang W, Li H, Fang S, Yang Z. Data-driven investigation of ship fuel consumption integrating causal inference and hierarchical analysis. *Transp Res D: Transp Environ* 2026;159:105491. <https://doi.org/10.1016/j.trd.2026.105491>.
- [6] Feng Y, Wang X, Cao Y, Zhou J, Wang J, Yang Z. Hierarchical relationships analysis: a novel data-driven framework to assess the influential factors of marine accidents. *Reliab Eng Syst Saf* 2027;277:113356. <https://doi.org/10.1016/j.res.2026.113356>.
- [7] Wang X, Zhao G, Cao L, Yu J, Li Y, Wang H, Cao W. Structural causality hierarchy analysis: a novel data-driven integrated method for assessing occupational accidents in maritime transportation. *Saf Sci* 2026;196:107086. <https://doi.org/10.1016/j.ssci.2025.107086>.
- [8] Wang X, Zhao G, Li T, Zhao J, Zhao X, Karatug Ç, Cao W. Context matters in just culture: translating formal just culture into lived near-miss reporting experience in maritime transportation. *Saf Sci* 2027;206:107518. <https://doi.org/10.1016/j.ssci.2026.107518>.
- [9] Cao L, Zhou M, Chen Y, Wang Q, Zhang Z, Li J, Wang X. Intelligent anchor-position detection algorithm based on multi-level spatial constraints. *Ocean Eng* 2026;362:126399. <https://doi.org/10.1016/j.oceaneng.2026.126399>.
- [10] International Maritime Organization. Resolution A.1155(32): procedures for port state control, 2021. International Maritime Organization; 2021.
- [11] Paris MoU on Port State Control. Overview of deficiency codes. Paris MoU; 2024.
- [12] Gard P&I Club. Engine room fires are still a major concern. Arendal, Norway: Gard; 2025.
- [13] Associated Press. The arrested captain of a cargo vessel involved in North Sea collision is Russian, ship's owner says. Associated Press; 2025.
- [14] Marine Accident Investigation Branch. Solong and Stena Immaculate interim report published. GOV.UK; 2025.
- [15] 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* 2026;266:111648. <https://doi.org/10.1016/j.res.2025.111648>.
- [16] 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. <https://doi.org/10.1016/j.res.2022.108934>.
- [17] Ma J, Cao L, Feng Y, Karatug Ç, Buber M, Wang X. Intelligent analysis of ship collision accidents via low-rank adaptation-based fine-tuning of medium-scale large language Models. *Reliab Eng Syst Saf* 2026;275:112774. <https://doi.org/10.1016/j.res.2026.112774>.
- [18] Fan L, Wang M, Yin J. The impacts of risk level based on PSC inspection deficiencies on ship accident consequences. *Res Transp Bus Manag* 2019;33:100464. <https://doi.org/10.1016/j.rtbm.2020.100464>.
- [19] Li T, Wang X, Zhang Z, Feng Y. A novel feature engineering method for severity prediction of marine accidents. *J Mar Eng Technol* 2025;1–16. <https://doi.org/10.1080/20464177.2025.2469336>.
- [20] Fan L, Zhang M, Yin J, Zhang J. Impacts of dynamic inspection records on port state control efficiency using Bayesian network analysis. *Reliab Eng Syst Saf* 2022;228:108753. <https://doi.org/10.1016/j.res.2022.108753>.
- [21] Fan L, Huang S, Wang M, Li KX. Dynamics of port state control and impact on ship safety: case studies of Tokyo and Indian Ocean memorandum of understanding. *Case Stud Transp Policy* 2025;21:101518. <https://doi.org/10.1016/j.cstp.2025.101518>.
- [22] Liu R, Zhang M, Yan R, Yang Z. Causal analysis of ship inspection data and maritime accidents through causal neural networks. *Reliab Eng Syst Saf* 2026;271:112255. <https://doi.org/10.1016/j.res.2026.112255>.
- [23] Cao Y, Wang X, Yang Z, Wang J, Wang H, Liu Z. Research in marine accidents: a bibliometric analysis, systematic review and future directions. *Ocean Eng* 2023;284:115048. <https://doi.org/10.1016/j.oceaneng.2023.115048>.
- [24] Feng Y, Wang X, Chen Q, Yang Z, Wang J, Li H, Xia G, Liu Z. Prediction of the severity of marine accidents using improved machine learning. *Transp Res E: Logist Transp Rev* 2024;188:103647. <https://doi.org/10.1016/j.tre.2024.103647>.
- [25] Liu Y, Zhou Y, Liu X, Dong F, Wang C, Wang Z. Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: a case study of cancer-staging data in biology. *Engineering* 2019;5:156–63. <https://doi.org/10.1016/j.eng.2018.11.018>.
- [26] Engelmann J, Lessmann S. Conditional wasserstein GAN-based oversampling of tabular data for imbalanced learning. *Expert Syst Appl* 2021;174:114582. <https://doi.org/10.1016/j.eswa.2021.114582>.
- [27] Kotelnikov A, Baranchuk D, Rubachev I, Babenko A. TabDDPM: Modelling tabular data with diffusion models. Krause A, Brunkill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research: PMLR; 2023. p. 17564–79.
- [28] Müller-Franzes G, Niehues JM, Khader F, Arasteh ST, Haaburger C, Kuhl C, Wang T, Han T, Nolte T, Nebelung S, Kather JN, Truhn D. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Sci Rep* 2023;13:12098. <https://doi.org/10.1038/s41598-023-39278-0>.
- [29] 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. <https://doi.org/10.1016/j.res.2021.107513>.
- [30] International Maritime Organization. Resolution MSC.255(84): adoption of the casualty investigation code. London: international maritime organization; 2008.
- [31] Cao Y, Wang X, Wang Y, Fan S, Wang H, Yang Z, Liu Z, Wang J, Shi R. Analysis of factors affecting the severity of marine accidents using a data-driven Bayesian network. *Ocean Eng* 2023;269:113563. <https://doi.org/10.1016/j.oceaneng.2022.113563>.
- [32] UNCTAD. Review of Maritime Transport 2024. In: United Nations conference on trade and development; 2024.
- [33] Yang Z, Yang Z, Yin J. Realising advanced risk-based port state control inspection using data-driven Bayesian networks. *Transp Res A: Policy Pract* 2018;110:38–56. <https://doi.org/10.1016/j.tra.2018.01.033>.
- [34] Yang Z, Yu Q, Yang Z, Wan C. A data-driven Bayesian model for evaluating the duration of detention of ships in PSC inspections. *Transp Res E: Logist Transp Rev* 2024;181:103371. <https://doi.org/10.1016/j.tre.2023.103371>.
- [35] Wang Di-ni, Li L, Zhao D. Corporate finance risk prediction based on LightGBM. *Inf Sci* 2022;602:259–68. <https://doi.org/10.1016/j.ins.2022.04.058>.
- [36] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 785–94. San Francisco, California, USA: Association for Computing Machinery.
- [37] Zhang Z, Jung C. GBDT-MO: Gradient-boosted decision trees for multiple outputs. *IEEE Trans Neural Netw Learn Syst* 2021;32:3156–67. <https://doi.org/10.1109/TNNLS.2020.3009776>.
- [38] Huang G, Wu L, Ma X, Zhang W, Fan J, Yu X, Zeng W, Zhou H. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *J Hydrol* 2019;574:1029–41. <https://doi.org/10.1016/j.jhydrol.2019.04.085>.
- [39] Shahraiki A, Abbasi M, Haugen Ø. Boosting algorithms for network intrusion detection: a comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. *Eng Appl Artif Intell* 2020;94:103770. <https://doi.org/10.1016/j.engappai.2020.103770>.
- [40] Speiser JL, Miller ME, Tootz J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst Appl* 2019;134:93–101. <https://doi.org/10.1016/j.eswa.2019.05.028>.
- [41] McDonnell K, Murphy F, Sheehan B, Masello L, Castignani G. Deep learning in insurance: accuracy and model interpretability using TabNet. *Expert Syst Appl* 2023;217:119543. <https://doi.org/10.1016/j.eswa.2023.119543>.
- [42] Tang J, Deng C, Huang GB. Extreme learning machine for multilayer perceptron. *IEEE Trans Neural Netw Learn Syst* 2016;27:809–21. <https://doi.org/10.1109/TNNLS.2015.2424995>.
- [43] Chauhan VK, Dahiya K, Sharma A. Problem formulations and solvers in linear SVM: a review. *Artif Intell Rev* 2019;52:803–55. <https://doi.org/10.1007/s10462-018-9614-6>.
- [44] Shi J, Liu Z. Accident data-driven consequence analysis in maritime industries. *J Mar Sci Eng* 2025;13(1):0117. <https://doi.org/10.3390/jmse13010117>.
- [45] Fan L, Zhao Z, Yin J. Research on dynamic influence mechanism of port state control and ship risk level. *Ocean Coast Manag* 2024;250:107028. <https://doi.org/10.1016/j.ocecoaman.2024.107028>.
- [46] Canbek G. BenchMetrics Prob: benchmarking of probabilistic error/loss performance evaluation instruments for binary classification problems. *Int J Mach Learn Cybern* 2023;14:3161–91. <https://doi.org/10.1007/s13042-023-01826-5>.
- [47] Yi H, Hou L, Jin Y, Saeed NA, Kandil A, Duan H. Time series diffusion method: a denoising diffusion probabilistic model for vibration signal generation. *Mech Syst Signal Process* 2024;216:111481. <https://doi.org/10.1016/j.ymssp.2024.111481>.
- [48] Linder J, Bogard N, Rosenberg AB, Seelig G. A generative neural network for maximizing fitness and diversity of synthetic DNA and protein sequences. *Cell Syst* 2020;11:49–62. <https://doi.org/10.1016/j.cels.2020.05.007>. e16.
- [49] Deng J, Wang S, Ye J, Ji L, Wang Y. DGRM: Diffusion-GAN recommendation model to alleviate the mode collapse problem in sparse environments. *Pattern Recognit* 2024;155:110692. <https://doi.org/10.1016/j.patcog.2024.110692>.
- [50] Azqadan E, Jahed H, Arami A. Predictive microstructure image generation using denoising diffusion probabilistic models. *Acta Mater* 2023;261:119406. <https://doi.org/10.1016/j.actamat.2023.119406>.
- [51] Garcia LPF, Lehmann J, de Carvalho A, Lorena AC. New label noise injection methods for the evaluation of noise filters. *Knowl-Based Syst* 2019;163:693–704. <https://doi.org/10.1016/j.knsys.2018.09.031>.
- [52] Staib M, Reddi S, Kale S, Kumar S, Sra S. Escaping saddle points with adaptive gradient methods. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research: PMLR; 2019. p. 5956–65.
- [53] Feng R, Zhao D, Zha Z-J. Understanding noise injection in GANs. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research: PMLR; 2021. p. 3284–93.
- [54] Orvieto A, Kersting H, Proske F, Bach F, Lucchi A. Anticorrelated noise injection for improved generalization. Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G, Sabato S, editors. In: Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research: PMLR; 2022. p. 17094–116.
- [55] Cao W, Wang X, Shu Y, Li H, Zhou J, Yang Z. An integrated method of advanced optimisation and adaptive ensemble learning for ship fuel consumption prediction.

- Transp Res C: Emerg Technol 2026;188:105659. <https://doi.org/10.1016/j.trc.2026.105659>.
- [56] Cao W, Wang X, Li J, Zhang Z, Cao Y, Feng Y. A novel integrated method for heterogeneity analysis of marine accidents involving different ship types. Ocean Eng 2024;312:119295. <https://doi.org/10.1016/j.oceaneng.2024.119295>.
- [57] Crestelo Moreno F, Roca Gonzalez J, Suardiaz Muro J, García Maza JA. Relationship between human factors and a safe performance of vessel traffic service operators: a systematic qualitative-based review in maritime safety. Saf Sci 2022;155:105892. <https://doi.org/10.1016/j.ssci.2022.105892>.
- [58] Adhita I, Fuchi M, Konishi T, Fujimoto S. Ship navigation from a Safety-II perspective: a case study of training-ship operation in coastal area. Reliab Eng Syst Saf 2023;234:109140. <https://doi.org/10.1016/j.res.2023.109140>.