Safety Science 204 (2026) 107384

Contents lists available at ScienceDirect

Safety Science

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

An automated Bayesian network framework for maritime cybersecurity risk assessment powered by large language models☆

Yunfeng Zhao [a], Huanhuan Li [b,*], Trung Thanh Nguyen [a], Christian Matthews [c], Zaili Yang [a,*]

[a] Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool, UK
[b] School of Engineering, University of Southampton, Southampton, UK
[c] Department of Maritime and Mechanical Engineering, Liverpool John Moores University, Liverpool, UK

ARTICLE INFO

Keywords:
Maritime cybersecurity
Large language model
Bayesian network
Risk influential factors
Probabilistic inference

ABSTRACT

The accelerating digitalisation of the maritime sector has increased its operational efficiency while simultaneously heightening its vulnerability to cyber-induced crises and emergency situations. Cyber incidents such as ransomware attacks, Global Positioning System (GPS) jamming, and phishing can trigger cascading system-level disruptions, undermining maritime operations, safety, and the resilience of critical transport infrastructures. To address this challenge, this study proposes an automated and scalable Bayesian Network (BN) framework powered by Large Language Models (LLMs) to support decision-relevant maritime cybersecurity risk analysis and assessment. A LLM-based text-mining pipeline is developed to reconstruct a comprehensive maritime cyber incident dataset (2001–2025) from unstructured reports. The proposed pipeline enables systematic identification and classification of key Risk Influential Factors (RIFs), which are subsequently validated through accuracy and consistency evaluations to ensure reliability and auditability. These RIFs are then modelled as variables within a BN model, facilitating probabilistic inference, characterization of interdependencies, and explicit analysis of risk propagation pathways associated with cyber-induced operational disruptions. Comparative experiments demonstrate that the proposed LLM-BN framework outperforms conventional BN models in terms of data consistency, inference accuracy, and structural interpretability. Beyond methodological contributions, the framework provides actionable insights to support emergency preparedness, risk prioritisation, and resilience enhancement for maritime operators, infrastructure managers, and regulators. By integrating structured knowledge extracted from textual sources with probabilistic risk modelling, this study contributes a decision-support framework for understanding, assessing, and managing cyber risks in complex socio-technical maritime systems, with broader applicability to other critical infrastructure sectors.

1. Introduction

Maritime transportation is a fundamental pillar of global trade, playing a critical role in international logistics networks and supply chain operations (Li et al., 2023; Potamos et al., 2024). The ongoing digital transformation of the maritime sector has accelerated the adoption of the Internet of Things (IoT) technologies, interconnected cyber-physical systems, automated, and real-time data exchange mechanisms. While these developments have significantly enhanced operational efficiency, they have simultaneously increased systemic exposure to cyber-induced vulnerabilities and disruptions (Nguyen et al., 2022; Yoo et al., 2021). Consequently, cyber incidents in maritime environments can no longer be regarded as isolated technical anomalies. Rather, they increasingly act as initiating events capable of triggering operational emergencies and cascading disruptions that propagate across complex and tightly coupled socio-technical systems, thereby posing significant challenges to maritime safety, resilience, and operational continuity.

Cyber threats such as ransomware attacks, Global Positioning System (GPS) jamming, phishing, and malware intrusions have been reported to disrupt vessel navigation, port operations, offshore installations, and logistics coordination, leading to operational failures, economic losses,

☆ This article is part of a special issue entitled: 'Crisis and Emergency Management' published in Safety Science.

* Corresponding authors.
E-mail addresses: Huanhuan.Li@soton.ac.uk (H. Li), Z.Yang@ljmu.ac.uk (Z. Yang).

https://doi.org/10.1016/j.ssci.2026.107384
Received 9 February 2026; Received in revised form 12 June 2026; Accepted 16 July 2026
Available online 22 July 2026
0925-7535/© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

and heightened safety risks (Benmalek, 2024). These cyber-induced disruptions pose significant challenges for crisis preparedness, emergency response, and resilience management in maritime transportation systems, where failures in one subsystem may rapidly escalate into broader system-level emergencies. Consequently, maritime cybersecurity has emerged as a critical concern for regulators, operators, and policymakers. International regulatory bodies, most notably the International Maritime Organization (IMO) through its International Safety Management (ISM) Code, together with industry organisations such as the Baltic and International Maritime Council (BIMCO), have explicitly recognised cybersecurity as an integral component of maritime safety and risk management (Peng et al., 2025). However, translating these high-level regulatory principles into effective, evidence-based, and decision-supportive risk analysis tools for crisis and emergency management remains a substantial challenge.

Recent academic efforts have advanced maritime cybersecurity analysis by classifying cyber threats and assessing their impacts across different maritime domains (Ben Farah et al., 2022; Clavijo Mesa et al., 2024). Several studies integrated maritime-specific factors and proposed risk analysis approaches, such as the Maritime Cyber Risk Assessment (MaCRA) framework (Tam and Jones, 2019) and Multi-Criteria Decision-Making (MCDM) methods (Tusher et al., 2022), to support both model-based and framework-driven cyber risk assessment in the maritime domain. Conventional risk analysis and assessment methods have also been adapted for maritime contexts. The commonly used methods include Failure Mode and Effects Analysis (FMEA) and Hazard and Operability Analysis (HAZOP) (Baybutt, 2015; Yang et al., 2008). Quantitative methods such as Fault Tree Analysis (FTA) (Henriques de Gusmão et al., 2018), Event Tree Analysis (ETA) (Ferdous et al., 2011), Evidential Reasoning (ER) (Zhang et al., 2016), and Bayesian Networks (BNs) (Zhang et al., 2013) have been widely applied to model and assess risk, particularly within the maritime safety analysis domain. Recognising the limitations of individual techniques, researchers have increasingly combined these methods into hybrid approaches to enable more comprehensive analysis and evaluation. For instance, Chang et al. (2021) employed the FMEA method in conjunction with ER and a Rule-Based BN (RBN) to quantify the risk levels of identified hazards. Uflaz et al. (2024) quantified potential cyber-attack risks in maritime transportation by integrating Dempster-Shafer (DS) theory with Failure Modes, Effects, and Criticality Analysis (FMECA) and RBN modelling. Park et al. (2023) adopted a BN-driven FMEA approach to assess cybersecurity risks in the maritime domain. In addition, Mohsendokht et al. (2024) employed BN in combination with objective data to analyse over a decade of maritime cybersecurity incidents, and generated valuable insights into recurring risk patterns and system vulnerabilities. These studies demonstrate the value of probabilistic modelling for capturing uncertainty and interdependencies in maritime cyber risk. Nevertheless, their practical applicability remains constrained by limited data scalability, dependence on manually curated inputs, and substantial reliance on expert elicitation during model construction and parameterisation.

Meanwhile, recent advances in Natural Language Processing (NLP), driven by Large Language Models (LLMs) such as ChatGPT and DeepSeek, have introduced a new paradigm for text mining. Unlike conventional NLP methods that require extensive annotation and preprocessing, LLMs offer strong generalisation capability, semantic understanding, and prompt-based few-shot or zero-shot learning for downstream tasks (Deng et al., 2025; Liu et al., 2023). These features enable LLMs to extract structured insights from unstructured narratives (Ahmadi et al., 2025; Wei et al., 2026), making them well-suited for processing maritime cybersecurity incident reports, which are often sparse, heterogeneous, and unstandardised.

Despite these methodological advances, three important limitations remain. First, existing maritime cyber risk assessment approaches lack the automation and scalability required for large-scale, data-driven risk analysis. Many studies continue to rely on expert judgement, which limits reproducibility and the effective use of diverse incident data sources. Second, although LLMs have been increasingly applied in security intelligence and risk analysis, their suitability for maritime cybersecurity text mining remains largely unexplored. Third, most BN-based maritime cybersecurity models are developed from relatively small datasets and manually constructed knowledge bases. While these models provide useful insights, they may not adequately capture the complexity of relationships embedded in large collections of cyber incident data, limiting their ability to represent risk dependencies and support informed decision-making.

To address these research limitations, this study proposes an integrated LLM-BN framework for maritime cybersecurity risk assessment. It systematically combines LLM-enabled text mining with probabilistic graphical modelling to establish a scalable and reproducible pipeline of cyber-risk analysis. Specifically, prompt-engineered LLMs are employed to automatically classify Risk Influential Factors (RIFs) from unstructured maritime cyber incident reports. The extracted RIFs are subsequently transformed into structured variables and incorporated into a data-driven BN model to support probabilistic inference, dependency analysis, and identification of key factors associated with cyber-induced maritime incidents. In addition, the effectiveness of LLM-based text mining is systematically evaluated. The accuracy and consistency of the LLM-extracted RIFs are validated through manual verification, ensuring auditability and transparency. Furthermore, comparisons with conventional BN-based approaches demonstrate the potential of the proposed framework to improve data consistency, inference accuracy, model interpretability, and scalability in maritime cyber risk assessment.

The main contributions of this study are threefold.

(1) An automated LLM-based pipeline is developed to identify and structure RIFs from maritime cybersecurity incident reports.
(2) An integrated LLM-BN framework is proposed to support maritime cyber risk assessment by linking LLM-based information extraction with BN modelling.
(3) The proposed framework is evaluated through comparative analysis against conventional BN-based approaches, demonstrating improvements in data consistency, inference accuracy, model interpretability, and scalability.

The remainder of this paper is structured as follows. Section 2 reviews existing literature on maritime cybersecurity risk analysis, outlining key methodological developments and research gaps. Section 3 presents the proposed LLM-BN methodology, detailing the automated classification of RIFs and their integration into the BN model. Section 4 presents and analyses the results obtained from the proposed methodology, highlighting the enhanced performance achieved by the proposed LLM-BN framework. Section 5 discusses the real-world implications for crisis and emergency management practice and policy. Finally, Section 6 summarises the key findings, acknowledges limitations, and suggests directions for future research.

2. Literature review

This section reviews existing research relevant to this study. Section 2.1 reviews existing maritime cybersecurity risk analysis methods and their limitations. Section 2.2 summarises recent advances in LLMs for cybersecurity-related text analysis. Section 2.3 synthesises these findings to identify key research gaps and motivate the proposed approach.

2.1. Maritime cybersecurity risk analysis: State of the art and limitations

Early research on maritime cybersecurity has predominantly adopted qualitative and conceptual approaches, focusing on vulnerability identification, threat classification, and high-level mitigation strategies. A substantial body of review and survey literature has examined maritime cybersecurity from complementary perspectives. Cyber threats and

vulnerabilities affecting shipboard systems, Global Navigation Satellite Systems (GNSS) (Ben Farah et al., 2022; Chaal et al., 2023), autonomous and remotely operated vessels (Zagan et al., 2022), and offshore platforms (Harish et al., 2025) have been systematically reviewed in terms of attack types, exposure points, and technological weaknesses (Yu et al., 2023). The consequences and impacts of cyber incidents on maritime operations, logistics performance, ports, and maritime supply chains, particularly in terms of disruption, safety degradation, and operational risk, have been analysed in several domain-focused reviews (Cheung et al., 2021; Clavijo Mesa et al., 2024). More recently, research strategies and methodological perspectives, including research trends, thematic evolution, and future directions in maritime cybersecurity, have been synthesised through bibliometric and meta-analytic studies (Peng et al., 2025). These studies have been instrumental in mapping the threat landscape and highlighting the safety and operational implications of maritime cyber incidents. Bibliometric evidence further confirms the rapid growth of this research field, while revealing a dominant reliance on descriptive analyses and qualitative assessment methods.

Several studies have explicitly linked cybersecurity to maritime safety and reliability. Bolbot et al. (2020) proposed a cyber-risk assessment method that associates cyber threats with hazardous operational states of ship systems, embedding cybersecurity considerations within traditional maritime safety assessment paradigms. Similarly, Androjna et al. (2020) demonstrated that cyberattacks targeting navigation-related systems, such as positioning, sensors, and communication, can directly compromise situational awareness and navigational safety. These studies reinforce the view that maritime cybersecurity acts as a precursor risk to navigational accidents and system failures. However, such approaches remain largely qualitative or semi-quantitative, offering limited capability for uncertainty representation or predictive risk reasoning.

To enhance analytical structure, a range of model-driven and decision-analytic frameworks have been proposed. Tam and Jones, (2019, 2018) introduced the MaCRA framework, incorporating vulnerability characteristics, exploitability, and attack rewards, and later extending it to autonomous shipping contexts. Other studies have adopted structured decision-making approaches, including MCDM and dynamic supply-chain-oriented cyber risk assessment (Schauer et al., 2019; Tusher et al., 2022). While these frameworks improve transparency and enable systematic comparison of cyber risk scenarios, they remain predominantly qualitative or semi-quantitative. In particular, they do not explicitly model uncertainty or structural dependencies among RIFs, limiting their ability to analyse risk propagation and support effective, rational risk mitigation strategies.

Recognising these limitations, recent research has increasingly adopted probabilistic approaches, particularly BNs, to model maritime cybersecurity risks. Hybrid frameworks combining BNs with FMEA (Park et al., 2023), ER (Chang et al., 2021), Dempster-Shafer theory (Uflaz et al., 2024), or Z-number (Aydin et al., 2025) representations have been proposed to enhance causal reasoning and uncertainty quantification. These studies demonstrate the suitability of probabilistic graphical models for representing interdependencies among cyber risk factors in complex maritime systems. Nevertheless, most BN-based models remain heavily dependent on expert elicitation for risk factor selection, network structure definition, and parameter estimation. This reliance introduces subjectivity, limits reproducibility, and constrains scalability.

The availability of structured cyber incident datasets has enabled more data-driven investigations. For example, Mohsendokht et al. (2024) analysed historical maritime cyber incidents to identify dependencies among RIFs and support empirical risk analysis. Despite these advances, the extraction and structuring of RIFs from unstructured incident narratives still rely predominantly on manual interpretation and expert judgement. This bottleneck restricts timeliness, hampers adaptability to emerging cyber threats, and limits the ability of probabilistic models to support continuous risk updating and effective decision-making.

2.2. Large language models in cybersecurity risk analysis

AI has been increasingly applied to cybersecurity to enhance threat detection, classification, and response automation. Recent systematic reviews highlight the rapid expansion of machine learning and deep learning techniques for intrusion detection, malware analysis (Hasanov et al., 2024), spam filtering, and cyber threat intelligence (Motlagh et al., 2024; Xu et al., 2025). In the maritime context, AI-based approaches have primarily been explored for enhancing cyber threat detection and operational monitoring, demonstrating their potential to improve situational awareness in complex and data-intensive environments (Miller et al., 2025). However, these applications are predominantly detection-oriented and do not explicitly address the transformation of cyber incident data into structured inputs for formal risk assessment.

LLMs represent a major advancement in cybersecurity-related text analytics due to their ability to perform high-quality semantic understanding under limited- or zero-labelled conditions. This capability is particularly relevant for cybersecurity domains, where incident reports, advisories, and disclosure documents are largely unstructured and sparsely annotated. Conventional text mining approaches typically require large volumes of labelled data to perform tasks such as Named Entity Recognition (NER), text classification, and topic modelling, which limits their applicability in safety- and security-critical domains. In contrast, LLMs enable effective information extraction and semantic interpretation even when labelled data are scarce (Mao et al., 2024).

Empirical studies show that LLM-based methods outperform traditional NLP techniques in entity extraction and text classification (Z. Wang et al., 2023), while prompt-based frameworks such as TopicGPT further demonstrate their flexibility for automated topic discovery in unstructured corpora (Pham et al., 2024). GPT-NER, which leverages LLMs for entity recognition (S. Wang et al., 2023). Moreover, domain-adapted and fine-tuned models, such as CySecBERT (Bayer et al., 2024), CyberLLaMA (H. Zhang et al., 2025), and SecLMNER (Y. Zhang et al., 2025), have demonstrated substantial performance improvements in cybersecurity-specific NER and information extraction tasks. LLMs have also been applied to cybersecurity text classification and threat identification problems, benefiting from their strong contextual reasoning capabilities (Ahmadi et al., 2025). Recent surveys of LLM applications in cybersecurity consistently report improvements in information extraction accuracy, contextual understanding, and cyber intelligence summarisation. Nevertheless, these reviews also highlight a key limitation: most LLM-based studies focus on extracting or classifying cyber-relevant entities and events in isolation. The extracted information is rarely structured in a way that supports downstream risk modelling, while relationships and dependencies among risk factors are often not explicitly represented.

Beyond general NLP tasks, LLMs have shown increasing potential in safety and risk-oriented applications. Prior studies have integrated LLMs with structured analytical frameworks to extract risk factors from accident reports in coal mining (Du and Chen, 2025), reliability assessment (Pang et al., 2024), and transportation systems (Zhen and Yang, 2025), achieving improved automation, consistency, and interpretability compared with manual processing. These findings indicate that LLMs can effectively transform heterogeneous textual evidence into structured, analysis-ready inputs for downstream risk modelling. In the maritime domain, applications of LLMs remain limited. Liu et al. (2026) proposed Pirate-GPT, a locally deployed LLM framework for offline anti-piracy decision support and knowledge retrieval. This work illustrates the feasibility of deploying LLMs in safety–critical maritime settings while emphasising reliability, robustness, and operational trust. However, such applications focus primarily on knowledge access and decision support rather than on quantitative cybersecurity risk modelling.

More broadly, LLMs have also been employed to enhance

cybersecurity analysis, including cyberattack scenario generation (Mudassar Yamin et al., 2024), threat detection in complex cyber-physical infrastructures such as satellite systems (Hassanin et al., 2025), and the construction of cyber threat intelligence knowledge graphs that reveal interconnections among vulnerabilities and attack patterns (Huang and Xiao, 2024). Despite these advances, the use of LLMs in maritime cybersecurity remains limited, particularly in systematically converting unstructured cyber incident narratives into structured variables suitable for formal risk assessment.

In summary, existing research demonstrates that LLMs are highly effective at automating the extraction and interpretation of cybersecurity-relevant information from unstructured text. However, current applications largely terminate at the information extraction or detection stage. The integration of LLM-derived knowledge with formal risk modelling frameworks, such as probabilistic graphical models, remains largely unexplored in the maritime cybersecurity domain. Addressing this gap is essential for enabling scalable, data-driven, and analytically rigorous maritime cybersecurity risk assessment.

2.3. Research gaps

Despite growing attention to maritime cybersecurity, current research remains fragmented across qualitative risk assessment, structured decision-analytic methods, probabilistic modelling, and data-driven AI techniques. Three important gaps continue to limit the development of robust and evidence-based maritime cyber risk assessment.

First, many existing approaches rely heavily on expert judgement and manually curated information, with limited use of evidence derived directly from cyber incident reports. As a result, risk assessments are often difficult to reproduce and validate against operational experience.

Second, BNs provide a useful framework for representing uncertainty and interdependencies in maritime cyber risk. However, the development of BN models remains highly dependent on manual knowledge elicitation and model construction, which limits their scalability and the use of large incident datasets.

Third, recent advances in LLMs have shown strong potential for extracting structured information from unstructured cybersecurity text. However, their application has largely focused on information extraction and threat detection, with limited integration into formal risk assessment methodologies. Methods for systematically incorporating LLM-derived knowledge into probabilistic risk models remain underdeveloped, particularly in the maritime cybersecurity domain.

3. Methodology

This section presents the methodological framework developed for this study, which integrates LLM with BN to achieve automated and data-driven maritime cybersecurity risk analysis. Section 3.1 introduces the proposed framework. Section 3.2 describes the processes of data collection and pre-processing, including the acquisition of incident reports, text cleaning, and data structuring to ensure high-quality inputs for model development. Section 3.3 details the LLM-based RIFs classification procedure, highlighting the prompt design, classification logic, and validation strategy to ensure accurate and consistent identification of RIFs. Section 3.4 presents the Tree-Augmented Naive (TAN)-based BN modelling, which captures dependencies among extracted RIFs to perform reasoning, inference, and risk assessment. Finally, Section 3.5 concludes with comprehensive model validation, covering predictive validation, mutual information, joint probability analysis, and the assessment of true risk influence, thereby demonstrating the robustness and practical applicability of the proposed approach.

3.1. The proposed framework

This study develops an integrated LLM-BN framework for maritime cybersecurity risk assessment. Maritime cyber incident reports are collected and analysed using prompt-engineered LLMs to identify and classify RIFs from unstructured narratives. The extracted factors are converted into structured variables and used to construct a BN model, which represents dependencies among risk factors and supports probabilistic inference and risk propagation analysis. The framework provides a systematic process for transforming incident reports into probabilistic risk models. The overall architecture of the framework is shown in Fig. 1.

3.2. Dataset construction

3.2.1. Data collection and pre-processing

Following a review of publicly available maritime cybersecurity incident repositories, the Maritime Cybersecurity Accident Database (MCAD), curated by the Maritime IT Security research group at NHL Stenden University of Applied Sciences, was selected as the primary data source for this study (Pijpker et al., 2024). MCAD provides a concise summary for each incident, including information on the affected country, attack type, and reported consequences, making it suitable for RIF identification and classification. Based on these incident summaries, RIFs were classified according to the definitions established in the RIF framework. To support this process, an automated data acquisition workflow was developed to retrieve, process, and structure MCAD records. The workflow consists of four stages:

1) Data retrieval: A web crawler was deployed to compile a comprehensive list of reported maritime cybersecurity incidents. Each incident entry was accessed via its unique identifier to obtain both detailed metadata and the full textual description of the event.
2) Basic factors extraction: For every record, the factor Reference Number, Year, Month, Title, and Summary were extracted. The first three were retained as temporal and administrative metadata, while the Title and Summary fields constituted the primary textual input for analysis.
3) Data validation: Redundant and incomplete entries were removed, and all remaining incident reports were cleaned to eliminate anomalous or non-printable characters, ensuring compatibility with subsequent LLM processing.
4) Corpus consolidation: The validated and structured data were converted into a CSV format, enabling compatibility with downstream NLP and probabilistic modelling tasks.

As of May 2025, a total of 355 verified maritime cybersecurity incident reports had been collected and refined following this workflow. This procedure produced a coherent and lexically standardised corpus, optimised for subsequent processing by the LLM-based extraction pipeline.

3.2.2. Dataset distribution and assessment

Before conducting RIF extraction and probabilistic modelling, it is necessary to examine the temporal characteristics and quality of the dataset to understand its representativeness and potential limitations. Fig. 2 presents the annual distribution of the 355 maritime cybersecurity incidents recorded between 2001 and 2025.

Incident frequencies remained low before 2015, with fewer than five cases reported in most years. From 2016 onwards, the number of reported incidents increased steadily, followed by a marked rise after 2020. The highest frequencies were observed in 2023 (81 incidents), 2024 (67 incidents), and 2022 (61 incidents). Approximately 70% of all incidents occurred during 2021–2024, which may reflect the increasing digitalisation of maritime operations as well as greater awareness and reporting of cybersecurity events. Although the temporal distribution is uneven, the dataset captures the development of maritime cyber threats over a 25-year period and provides a substantial empirical basis for analysis.

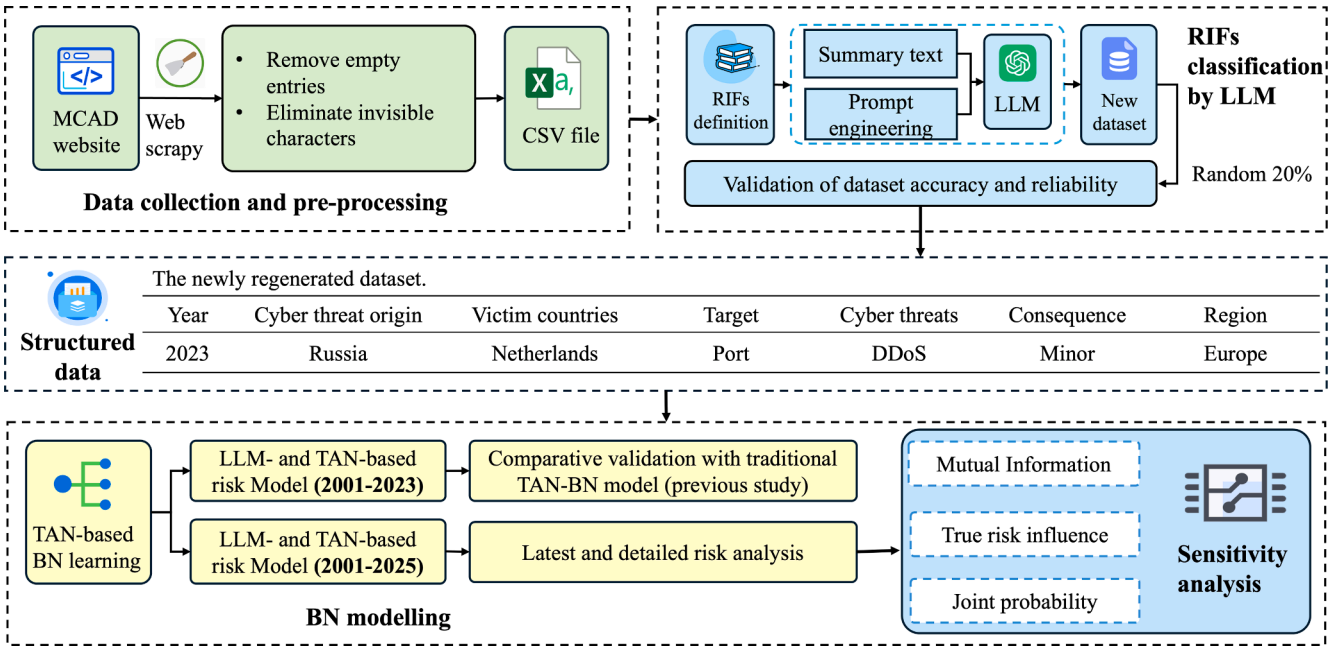

Fig. 1. Overview of research methodology.

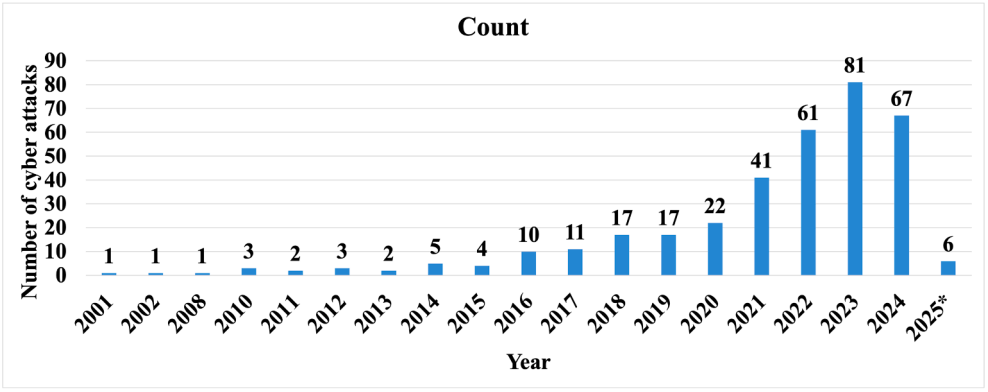

Fig. 2. Annual distribution of maritime cybersecurity incidents (2001-May 2025).

In addition to understanding the temporal distribution of incidents, the suitability of MCAD for RIF extraction was assessed through a manual review of randomly selected records. The incident summaries were found to contain information covering several key dimensions of maritime cyber incidents, including threat origin, affected country or region, target entity, attack type, and operational consequences. This level of coverage provides sufficient contextual detail to support the identification and classification of RIFs. Consequently, the summary field was adopted as the primary source of information for the subsequent LLM-based extraction process.

Despite its broad coverage, several limitations of MCAD should be acknowledged. As with other publicly available incident databases, the dataset is restricted to reported and documented events and may not capture incidents that remain undisclosed or confidential. Consequently, the distribution of incidents may be influenced by differences in reporting practices, regulatory transparency, and media attention across regions and sectors. In addition, the level of detail provided in incident summaries varies between records. While most summaries contain sufficient information to support RIF identification, some reports provide only limited contextual information regarding attack mechanisms, contributing factors, or organisational conditions. These limitations may affect the representation of certain risk factors and should be considered when interpreting the results of the subsequent BN analysis. Nevertheless, MCAD remains the most comprehensive publicly available source of maritime cybersecurity incident data and provides a valuable empirical basis for investigating cyber-risk relationships in maritime socio-technical systems (Badea et al., 2025; Hanafiah et al., 2025).

3.3. *Automated classification of RIFs using LLM*

3.3.1. *RIFs definition*

Following the compilation of maritime cybersecurity incident reports, the subsequent stage involves classifying RIFs, which form the structured input variables for the BN risk model. Clearly defining these factors is a critical prerequisite for ensuring both the accuracy of automated extraction and the interpretability of the resulting probabilistic model. To enable benchmark comparability with established studies, the RIFs adopted in this research largely align with the factor taxonomy proposed by Mohsendokht et al. (2024), while introducing minor modifications to improve clarity and adaptability to automated extraction. The final RIFs taxonomy, which serves as the foundation for both LLM-based extraction and BN modelling, is presented in Table 1. Each factor definition includes its conceptual scope, representative examples, and textual identification criteria. This structured schema provides the

Table 1
States of RIFs.

RIFs	Definition	States
Year	Year of incident.	2001–2025
Cyber threat origin	Country of origin of the cyberattack.	Russia, China, North Korea, Iran, Unknown. Values below 8 will be categorised as "Other".
Victim countries	Country where the cyberattack occurred.	Italy, France, Australia, Greece, Norway, South Korea, United States, Canada, Netherlands, Germany, United Kingdom, Russia, China, Unknown. Values below 9 will be categorised as "Other".
Region	Region (continent) where the cyberattack occurred.	East Asia, Europe, Middle east & North Africa, North America. Values below 37 will be categorised as "Other".
Target	Type of target of the cyberattack.	Port, Vessel, Offshore structure, Shipping company, Shipbuilding firm, Energy distributor. Values below 14 will be categorised as "Other".
Cyber threats	Cyberattack method.	Ransomware, Spoofing, Hacking, Malware, DDoS, Jamming, Phishing.
Consequence	Consequence severity of the cyberattack.	Based on BIMCO (Guidelines on cyber security onboard ships, 2024): Major: Incidents resulting in significant operational disruption, prolonged service outage, substantial data breaches, major financial losses, or adverse impacts on critical maritime infrastructure and business operations. Minor: Incidents resulting in limited or temporary disruption, isolated unauthorised access, brief service interruption, or signal manipulation without evidence of significant operational, financial, or safety consequences.

conceptual and operational bridge between unstructured textual narratives and the quantitative reasoning capabilities of the BN model, ensuring the resulting risk analysis remains both data-driven and semantically interpretable.

To ensure reliable parameter estimation and avoid data sparsity in the BN, low-frequency states were aggregated into an "Other" category when the number of observations was insufficient for robust probabilistic learning. This discretisation strategy follows common practice in data-driven BN modelling and aims to balance model interpretability, statistical robustness, and information retention.

3.3.2. Prompt engineering

Prompt engineering refers to the systematic design and optimisation of input instructions that guide LLMs to perform downstream tasks such as text classification, entity extraction, and structured information retrieval. Unlike traditional supervised NLP pipelines, which require large labelled datasets, feature engineering, and task-specific model training, prompt engineering leverages pretrained LLMs to adapt to new tasks via natural-language instructions (Liu et al., 2023; Sahoo et al., 2025). This makes it particularly suitable for unstructured, narrative-rich domains such as maritime cybersecurity incident reporting, where labelled data are scarce and textual formats vary widely.

Formally, let M_θ denote an LLM parameterised by pretrained weights θ. A prompt P is designed as a structured instruction sequence containing task directives, domain definitions, and a few optional few-shot examples:

$$P = (I, C, E) \tag{1}$$

where I denotes explicit task instructions, C embeds domain knowledge or constraints, E provides example input–output demonstrations (few-shot prompting). Given an unstructured text sample x, such as an incident summary, the model generates an output y according to:

$$y = \underset{y'}{\arg\max} p_\theta(y'|x, P) \tag{2}$$

$p_\theta(y'|x, P)$ is the conditional probability assigned by the LLM parameterised by θ.

Prompt engineering seeks to optimise the prompt structure P such that domain-specific knowledge and classification criteria are explicitly incorporated into the inference process, thereby improving classification accuracy, consistency, and reproducibility. A well-designed prompt reduces ambiguity, constrains the model's reasoning process, and ensures that generated outputs reflect explicit textual evidence rather than unsupported inference.

To enhance the reliability and transparency of LLM outputs in text-mining tasks, this study incorporates chain-of-thought (CoT) prompting, which instructs the model to articulate its reasoning process before generating a final classification label (Wei et al., 2022). Rather than outputting a category directly, CoT prompting guides the model to identify relevant textual evidence, explain how this evidence corresponds to predefined categories, and disregard irrelevant or ambiguous content. This step-by-step reasoning improves interpretability, reduces classification errors arising from implicit assumptions, and provides an inherent validation mechanism. Moreover, CoT outputs enable researchers to observe systematic misinterpretation patterns during development, supporting iterative refinement of the prompt design. Additionally, to further ensure consistency, machine-readability, and seamless integration with downstream analytical pipelines, the LLM is instructed to generate outputs in a strict JavaScript Object Notation (JSON) format. JSON provides a standardised schema that minimises variation across outputs, enables fully automated parsing for large-scale text-mining tasks, and facilitates error detection by flagging deviations from the expected structure. Fig. 3 illustrates an example of how the proposed framework classifies the severity of consequences based on the designed prompt structure.

3.3.3. Validation of text classification

Following prompt development, the reliability of the LLM-based classification process was evaluated to ensure the validity of the extracted data and mitigate potential errors arising from model hallucinations. The evaluation focused on two complementary dimensions: accuracy and consistency. Given the absence of a large annotated corpus, 20% of the dataset was manually labelled to establish a gold-standard validation set. To assess the stability of model outputs, the classification process was repeated ten times under identical conditions. The combination of manual validation and repeated evaluation provides a robust assessment of classification reliability and reproducibility for small- to medium-scale text-mining applications.

(1) Comparative evaluation of LLMs.

To further evaluate the robustness and generalisability of the proposed extraction framework, additional comparative experiments were conducted using multiple LLMs, including GPT-4o, Gemini-2.5-Flash, and DeepSeek-V3.2. These models represent different architectural and deployment paradigms, ranging from proprietary commercial models to open-source-oriented alternatives. To ensure comparability, all models were evaluated using identical prompts, classification criteria, input data, and extraction procedures. The temperature parameter was fixed at zero to minimise stochastic variation and ensure deterministic outputs.

All models were evaluated using the same prompt structure, classification criteria, and validation subset comprising 20% of the dataset

```
Task: Analyze the cybersecurity incident report and classify it by consequence severity.

Severity categories:
- Major: The attack causes substantial disruption to the targeted entity's operations, resulting in significant physical or financial damage, or involving serious data theft and credential compromise.
- Example: The NotPetya ransomware attack on Maersk, which caused a $300 million loss.
- Example: The hacking of India's Jawaharlal Nehru Port Container Terminal in February 2022, which caused a five-day shutdown.
- Minor: The attack has less severe consequences, such as being successfully thwarted or quickly recovered from, with limited or no lasting damage.

Output format:
{{"result": "Major or Minor", "reason": "Explanation"}}

Report: "{incident_report}"
```

Fig. 3. Prompt design for classifying consequence severity.

(71 reports) (Gholamy et al., 2018). To ensure comparability, identical input settings and extraction procedures were applied across all experiments. Classification performance was assessed using Overall Accuracy, Macro Precision, Macro Recall, and Macro F1-score, which are widely adopted metrics in text classification and information extraction studies (Powers and Ailab, 2011; Sokolova and Lapalme, 2009).

Overall Accuracy was used to measure the proportion of correctly classified instances across the entire validation set. Given the imbalanced distribution of cyber threat categories within the dataset, Macro Precision, Macro Recall, and Macro F1-score were also adopted to provide a balanced evaluation of classification performance by assigning equal importance to each category regardless of its frequency. This evaluation strategy enables a systematic assessment of both overall classification correctness and category-level extraction performance.

Traditional supervised NLP approaches were not adopted as baselines because the maritime cybersecurity domain currently lacks sufficiently large and standardised labelled corpora required for reliable model training and fine-tuning. In contrast, prompt-based LLM approaches are better suited to small-scale and heterogeneous maritime incident datasets due to their strong zero-shot and few-shot generalisation capabilities.

The comparative evaluation enables systematic assessment of the trade-offs among extraction performance, model accessibility, and practical deployment in maritime cybersecurity text-mining applications.

(2) Consistency evaluation of the selected LLM.

The best-performing model identified in the comparative evaluation was subsequently subjected to a consistency assessment to examine the stability and reproducibility of the extracted results prior to BN construction.

The selected model was executed ten times using identical prompts, incident reports, and classification criteria. Repeated-run evaluation has been widely adopted to assess the stability and reproducibility of LLM-generated outputs in text annotation and classification tasks (X. Wang et al., 2023). Consistency was quantified as the proportion of identical classifications obtained across the ten repeated runs. Consistency scores were calculated for each target field and subsequently aggregated to derive an Overall Consistency score. The temperature parameter was fixed at 0.2 to introduce a modest degree of stochasticity while still maintaining relatively consistent outputs, allowing the stability of the prompt design to be evaluated.

A higher consistency value indicates that the prompt design, factor definitions, and classification schema are sufficiently stable and unambiguous. By extending the evaluation to ten repeated runs and reporting both field-level and overall consistency scores, the analysis provides stronger evidence regarding the robustness and reproducibility of the proposed LLM-based extraction framework, thereby increasing confidence in the reliability of the structured data used for subsequent BN construction.

3.4. TAN-based BN modelling

BNs have emerged as a robust probabilistic framework for risk assessment, providing powerful capabilities for both dependency analysis and data-driven inference. To enhance objectivity, reproducibility, and data dependence, this study adopts a purely data-driven modelling strategy by employing the TAN algorithm for BN structure learning. Unlike conventional BN approaches, which typically assume independence among predictor variables conditional on the target node, TAN introduces an additional dependency layer among features while maintaining a single parent–child relationship with the class node (Li et al., 2024; Mohsendokht et al., 2026). This improvement enables the model to capture inter-feature dependency patterns that may reflect meaningful real-world operational associations within maritime cybersecurity incident data. Consequently, TAN extends the representational capacity of conventional Naive Bayes while preserving tractable computational efficiency.

In this study, the cyber threat node was defined as the central node of the network, while the extracted RIFs were represented as explanatory variables. The TAN algorithm was applied to learn the network structure from the classified incident dataset, enabling the identification of dependencies among RIFs associated with different cyber threats. The resulting BN supports probabilistic inference, dependency analysis, and the exploration of cyber-risk propagation pathways.

Model development was conducted using Netica (version 5.12; Norsys Software Corp), which was used for both structure learning and parameter estimation. The learned network was subsequently employed for probabilistic analysis of relationships between cyber threats and associated risk factors.

3.5. Model validation

3.5.1. Predictive validation

The extraction framework was independently evaluated through manual verification, cross-model comparison, and consistency analysis prior to BN construction. Therefore, the objective of the predictive validation is not to provide a direct assessment of LLM extraction accuracy, but rather to examine whether the extracted structured dataset

preserves sufficient information to support downstream risk modelling and inference. If the extracted risk factors accurately capture the relationships described in maritime cyber incident reports, the resulting BN should demonstrate satisfactory predictive performance on unseen cases. To ensure a like-to-like comparison, data from 2001 to 2023 were extracted and compared with the findings in Mohsendokht et al. (2024) to demonstrate the accuracy of the LLM-based BN model. In contrast, this study employs the same dataset. However, all factors are extracted and constructed from unstructured incident reports using an LLM-based approach. To verify the predictive capability of the TAN-BN model, the dataset was divided into 15% and 85% according to the proportion of cyber threats, in order to maintain consistency with the number of test samples used in previous studies.

In binary classification, Precision, Recall, F1-measure, and Specificity are standard evaluation metrics derived from the confusion matrix. Precision measures the proportion of predicted positive instances that are truly positive, defined as

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall measures the proportion of actual positive instances that are correctly identified, given by

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

F1-measure is the harmonic mean of Precision and Recall, balancing both metrics, and is defined as

$$F1 = \frac{2 \times Precision \times Recall}{Presision + Recall} \tag{5}$$

Specificity measures the proportion of actual negative instances that are correctly classified, and is given by

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively.

3.5.2. Mutual information

Mutual information (MI) quantifies the statistical dependence between two random variables by measuring how much knowing one variable reduces uncertainty about the other. In other words, it captures the amount of shared information between variables and is zero if and only if the variables are independent. Unlike correlation, mutual information can detect both linear and nonlinear relationships. Formally, for two random variables X and Y, mutual information is defined as

$$I(x;y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{7}$$

where $p(x,y)$ is the joint distribution and $p(x), p(y)$ are the marginal distributions. In the BN model, mutual information is commonly used to assess the strength of dependency between variables and to guide structure learning by identifying pairs of nodes that are likely to be directly connected.

3.5.3. Joint probability

The joint probability table represents the likelihood of every possible combination of the class variable and attribute values, reflecting the dependencies defined by the network structure. It illustrates how the value of one variable influences another when a connection exists, thereby capturing both class-attribute and inter-attribute relationships. By examining the combined effects across temporal, spatial, and categorical dimensions, this analysis clarifies the complex evolution of cyber threats over time.

Formally, given a set of random variables $X_1, X_2, \cdots, X_n$, a BN model factorizes their joint probability distribution as

$$P(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)) \tag{8}$$

where $Pa(X_i)$ denotes the set of parent variables of X_i in the directed acyclic graph. This factorization is the core property of BN models, allowing complex high-dimensional joint distributions to be represented compactly using local conditional probability distributions.

3.5.4. True risk influence

The derivation of True Risk Influence (TRI) values builds upon the distinction between the highest and lowest risk contributions of each RIF. As shown in the joint probability table, the bolded figures represent the High-Risk Inference (HRI), where the probability of a voyage segment reaches its maximum when a particular RIF state is set to certainty. Similarly, the bolded figures also denote the Low Risk Inference (LRI), capturing scenarios of minimal impact. TRI is then calculated as the average of these two extremes, providing a balanced measure of the true influence exerted by each RIF on the associated voyage segment.

$$TRI = \frac{HRI + LRI}{2} \tag{9}$$

4. Results

This section presents and analyses the results obtained from the proposed methodology. It begins with the evaluation of LLM-based text extraction validation results, assessing the accuracy and reliability of the extracted information in section 4.1. Section 4.2 reports the outcomes of the TAN-based modelling, highlighting how dependencies among RIFs are learned and represented. Section 4.3 presents comprehensive model validation results, including predictive validation to evaluate model performance, MI analysis to quantify relationships between variables, joint probability assessment to capture combined risk behavior, and TRI analysis to identify the most impactful risk factors. Together, these results demonstrate the effectiveness and interpretability of the proposed modelling approach.

4.1. Text extraction validation result

Based on the validation procedure described in Section 3.3.3, the performance of different LLMs was systematically evaluated using the predefined validation subset and classification metrics. Table 2 presents the comparative performance of the three LLMs. GPT-4o achieved the best results across all evaluation metrics, with an Overall Accuracy of 0.9601 and a Macro F1-score of 0.9698, outperforming DeepSeek-V3.2 (0.9061) and Gemini-2.5-Flash (0.8662). The higher Macro Precision and Macro Recall obtained by GPT-4o indicate that it provided more balanced classification performance across different cyber threat categories, including less frequent classes. These results demonstrate that GPT-4o was more effective in extracting and classifying information from maritime cyber incident reports. Therefore, GPT-4o was selected for the subsequent consistency evaluation.

Based on the consistency evaluation procedure described in Section 3.3.3, GPT-4o, which achieved the best overall classification performance, was further assessed through ten repeated extraction runs to

Table 2
Comparative classification performance of different LLMs.

Model	Overall accuracy	Macro precision	Macro recall	Macro F1-score
GPT-4o	0.9601	0.9707	0.9710	0.9698
DeepSeek-V3.2	0.9061	0.9120	0.9143	0.9124
Gemini-2.5-Flash	0.8662	0.8785	0.8850	0.8803

evaluate the reproducibility of the proposed framework. Table 3 presents the consistency evaluation results of GPT-4o across ten independent runs. The model produced highly stable classification results, with an overall consistency of 97.42%, while field-level consistency ranged from 94.37% to 100.00%. Cyber threats achieved perfect consistency, and all remaining fields exceeded 94%. Only minor variations were observed for Region, which may be associated with differences in the interpretation of geographically related descriptions. The consistently high agreement across repeated runs indicates that the extraction framework generated reproducible outputs and provides confidence in the reliability of the structured data used for subsequent BN modelling.

4.2. TAN-based BN modelling

Following the methodological framework outlined in Section 3.4, a TAN-structured BN was developed to model the probabilistic relationships among the identified risk factors. To ensure model robustness, the input data were preprocessed through feature selection and normalisation procedures to reduce redundancy and enhance learning efficiency. The final TAN-structured BN, as illustrated in Fig. 4, visualises the learned dependencies among key variables, where directed arcs denote conditional influence relationships. The structure provides a clear overview of how RIFs interact to shape the overall risk profile.

4.3. Model validation

4.3.1. Predictive validation

As a result, the LLM-based BN model built using the training set achieved an accuracy exceeding 96%, outperforming the previous study (93.3%) (Mohsendokht et al., 2024). These results indicate that the structured dataset generated by the proposed extraction framework provides an effective basis for BN learning and inference. The improved predictive performance suggests that the extracted risk factors preserve meaningful relationships among maritime cyber risk variables while substantially reducing the manual effort required for dataset construction. Table 4 presents the confusion matrix comparing the prediction accuracy of the BN models.

In terms of detailed performance metrics, the current model also showed higher precision (0.95), recall (0.95), F1-measure (0.94), and specificity (0.99) compared with the previous model's precision (0.88), recall (0.92), F1-measure (0.89), and specificity (0.97). These results suggest that the current model not only produces more accurate and consistent predictions but also better distinguishes between different attack types while minimising false positives. Table 5 presents the performance metrics of predicted results based on the confusion matrix.

4.3.2. Mutual information

Table 6 presents the MI values for all nodes relative to the Cyber threats variable. As expected, the Cyber threats node itself exhibits the maximum MI value of 2.0925, establishing a baseline of 100% contribution. The Year variable demonstrates a notable impact, accounting for 16.7% of the MI relative to Cyber threats. This suggests that temporal patterns play a significant role in shaping the cyber threat landscape, as variations in the year of occurrence influence the likelihood distributions of different threat categories, reflecting the temporal evolution of attack techniques and target preferences.

The Cyber threat origin variable contributes 8% of the mutual information, indicating a moderate role in distinguishing among different cyber threat types. This finding suggests that the geographical or organisational source of a cyberattack provides meaningful insight into the likely nature of the threat, although its influence is less pronounced than that of temporal factors. Variables such as Consequence and Victim countries contribute 3.78% and 3.76%, respectively. While these variables bear some predictive significance, their influence on the immediate classification of threat types is comparatively lower. Attributes such as Target and Region show even smaller percentages, at 3.64% and 1.67%, respectively, indicating that although operational sectors and broader regional affiliations provide contextual information, they have a limited direct impact on the classification of threat types. The variance of beliefs associated with each node further substantiates these findings, offering additional insight into the uncertainty inherent in each variable's influence. Higher variance values indicate greater variability in belief updating when evidence is incorporated, reinforcing the mutual information outcomes.

Overall, the MI values confirm that temporal information, Year, Cyber threat origin, and Consequence are the most critical auxiliary variables influencing cyber threat classification. These insights provide empirical support for prioritising feature selection and evidence collection in future refinements of the model. By systematically quantifying the contribution of each node, the MI values analysis enhances the transparency and interpretability of the LLM-driven BN framework.

4.3.3. Joint probability

Table 7 presents the joint probability table of the current BN model. In the early years, incidents such as DDoS and hacking attacks registered a higher probability, peaking notably in 2001 and 2011. These peaks reflect the disruptive potential and novelty of these attack types at the time. However, a significant shift is observed after 2014, as spoofing, malware, and later ransomware incidents become more prevalent, mirroring the increasing digitalisation of critical infrastructures. The surge in ransomware attacks from 2020 onwards is particularly striking, culminating in unprecedented probability levels in 2022 and 2023. This highlights a recalibration of threat priorities within the maritime sector.

Regarding the origin of cyber threats, Russia stands out for its overwhelmingly high probability in ransomware attacks, far exceeding other countries across all threat categories. In contrast, countries like North Korea, China, and Iran exhibit more balanced threat profiles, with multiple cyber-attack types maintaining significant probability without a single type dominating. Interestingly, incidents attributed to "Unknown" sources consistently show high ransomware probability, underscoring the increasing sophistication and anonymisation tactics employed by modern attackers.

Region analysis confirms the concentration of high probability Ransomware incidents in Europe and North America, regions with dense maritime operations and advanced digital systems. In East Asia and the Middle East & North Africa, in addition to ransomware attacks, spoofing attacks also account for a significant proportion, indicating a strong association with key regional maritime hubs. Victim countries display that major maritime economies such as the USA, Germany, and France exhibit elevated ransomware probability, reflecting their substantial digital asset density and attractiveness to cybercriminals. The "Other" category, encompassing smaller or less digitally advanced states, shows disproportionately high ransomware probability, suggesting weaker cyber resilience and delayed response capabilities.

Target-specific patterns reveal that shipping companies and ports are particularly vulnerable to ransomware, each registering a probability above 50%. Shipbuilding firms and offshore structures also face significant exposure, though their threat profiles are more diversified, with notable contributions from Hacking and Malware. This suggests that operational settings closer to terrestrial infrastructure attract a broader range of threats, while mobile maritime assets increasingly face

Table 3
Consistency scores of GPT-4o over ten repeated runs.

Target field	Consistent classifications	Consistency (%)
Cyber threat origin	70/71	98.59
Victim countries	69/71	97.18
Region	67/71	94.37
Target	69/71	97.18
Cyber threats	71/71	100.00
Consequence	69/71	97.18
Overall Consistency	–	97.42

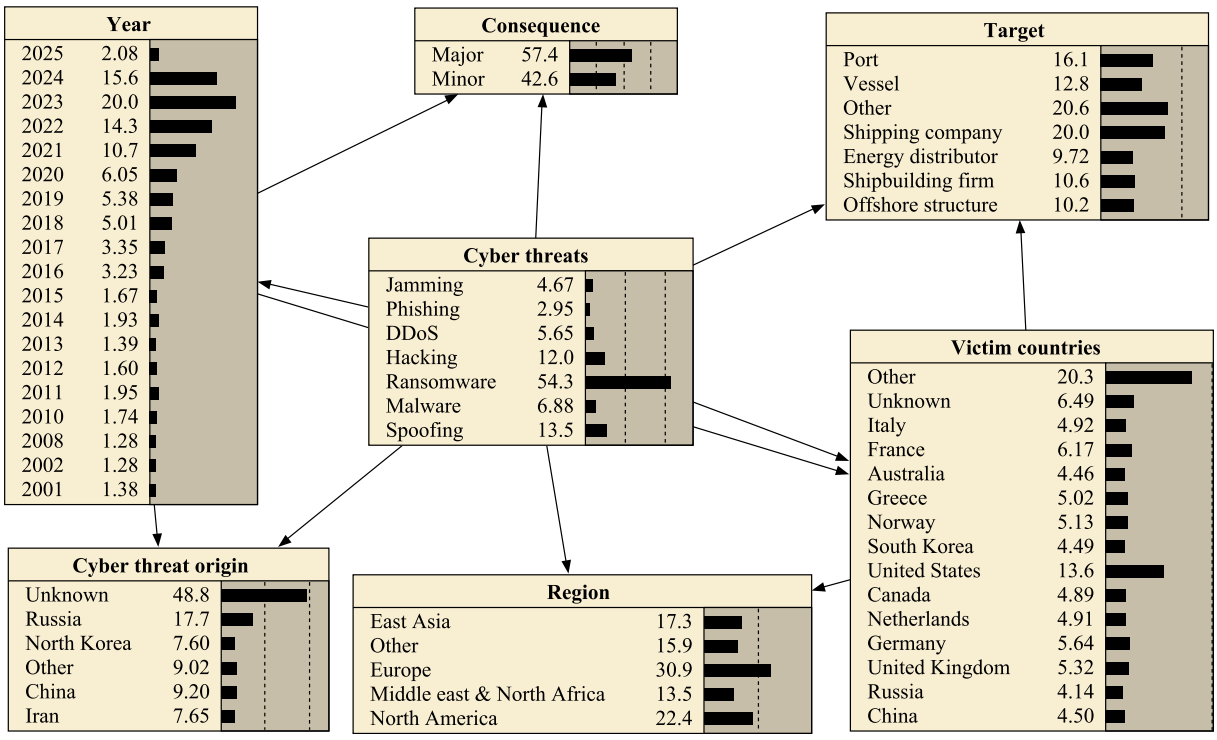

Fig. 4. BN Modelling.

Table 4
Confusion matrix of predicted results.

Actual predicted	Ransomware	DDoS	Hacking	Spoofing	Phishing	Jamming	Malware	Actual total	Accuracy rate (%)
Ransomware	17	0	0	0	0	0	0	17	100
DDoS	0	1	0	0	0	0	0	1	100
Hacking	0	0	2	0	0	0	1	3	66.7
Spoofing	0	0	0	4	0	0	0	4	100
Phishing	0	0	0	0	1	0	0	1	100
Jamming	0	0	0	0	0	1	0	1	100
Malware	0	0	0	0	0	0	2	2	100
Total	17	1	2	4	1	1	3	29	96.6

Table 5
The performance metrics of predicted results.

	Precision	Recall	F1-measure	Specificity
Ransomware	1	1	1	1
DDoS	1	1	1	1
Hacking	1	0.67	0.8	1
Spoofing	1	1	1	1
Phishing	1	1	1	1
Jamming	1	1	1	1
Malware	0.67	1	0.8	0.96
Average	0.95	0.95	0.94	0.99

Table 6
Mutual information between cyber threats and other RIFs.

Node	Mutual Info	Percentage	Variance of Beliefs
Cyber threats	2.0925	100	0.4878
Year	0.3489	16.7	0.0386
Cyber threat origin	0.1674	8	0.0241
Consequence	0.0790	3.78	0.0111
Victim countries	0.0787	3.76	0.0137
Target	0.0761	3.64	0.0122
Region	0.0350	1.67	0.0054

financially motivated attacks.

Analysis of the consequences of cyber incidents further underscores the evolving risk landscape. Incidents classified as “major” almost exclusively involve ransomware attacks, with probability exceeding 60%, highlighting their severe operational and financial implications. While “minor” incidents exhibit a more diverse threat profile, ransomware still maintains a dominant presence, underscoring its pervasive impact across all severity levels.

Overall, this joint probability analysis highlights a clear shift towards financially motivated, high-impact cyber threats, particularly Ransomware. While traditional threats such as DDoS and Hacking remain relevant, their influence has declined over time. These findings underscore the need for a context-sensitive approach to cyber risk assessment, focusing on resilience against emerging threat vectors while maintaining vigilance across the full spectrum of cyber risks.

4.3.4. *True risk influence*

Table 8 presents the TRI results obtained using the methodology described in Section 3.5.4. Ransomware exerts an overwhelming influence, achieving the highest TRI values in every category. Its dominance is particularly evident under the “Year” dimension, where the calculated TRI reaches 37.81, and remains consistently elevated across “Cyber threat origin” and “Victim countries”. This indicates that Ransomware

Table 7
The joint probability.

	Jamming	Phishing	DDoS	Hacking	Ransomware	Malware	Spoofing
Original	4.67	2.95	5.65	12	54.3	6.88	13.5
Year							
2001	9.15	7.12	**30**	13	16.5	10.8	13.4
2002	9.83	7.66	10.7	28	17.7	11.7	14.4
2008	9.83	7.66	10.7	28	17.7	11.7	14.4
2010	14.5	5.65	7.93	31	13.1	17.2	10.6
2011	12.9	5.04	7.07	**46.1**	11.7	7.67	9.5
2012	23.6	12.3	8.59	11.2	14.2	18.6	11.5
2013	9.07	7.06	19.8	12.9	16.3	21.5	13.3
2014	6.53	10.2	7.14	27.9	11.8	7.74	28.7
2015	7.58	**17.7**	8.28	21.6	13.6	8.98	22.2
2016	15.6	3.04	8.54	27.8	7.04	9.26	28.7
2017	3.76	8.79	4.11	32.2	**6.78**	**22.3**	22.1
2018	17.6	1.96	5.51	14.4	22.7	12	25.9
2019	2.35	3.65	2.56	**3.34**	25.3	11.1	**51.6**
2020	2.09	1.62	2.28	11.9	56.3	7.42	18.4
2021	1.18	1.83	**1.28**	11.7	63.5	8.36	12.1
2022	1.76	2.06	3.85	5.01	**82.4**	1.04	**3.87**
2023	**0.63**	**0.49**	6.19	7.18	78.3	**0.75**	6.47
2024	1.62	0.63	5.3	5.76	72.8	6.71	7.12
2025	**24.3**	9.46	13.3	25.9	10.9	7.2	8.91
Cyber threat origin							
Unknown	**2.5**	**1.28**	**2.88**	**6.8**	**74.6**	**3.26**	**8.64**
Russia	3.44	2.44	7.53	7.79	56.3	6.7	15.9
North Korea	**11.4**	5.68	9.04	**24**	**22.2**	10.9	16.8
Other	8.38	**6.01**	**9.44**	22	23.4	11.1	18.7
China	6.61	5.3	7.47	22.3	25.1	**15.7**	17.5
Iran	7.95	5.63	8.98	19.3	23.5	10.8	**23.9**
Region							
East Asia	6.69	3.36	5.17	14.2	44.1	8.12	**18.3**
Other	4.96	3.84	6.35	13.5	48.7	7.1	15.6
Europe	**2.7**	**2.19**	**4.73**	**8.57**	**64.2**	**5.93**	11.7
Middle East & North Africa	**8.18**	**4.11**	**7.88**	**14.6**	**38.7**	**8.57**	18
North America	3.51	2.35	5.46	12.5	61.9	6.06	**8.16**
Victim countries							
Other	2.88	**1.18**	3.2	**6.07**	**73.3**	4.01	9.37
Unknown	4.93	3.85	5.57	12.9	51.6	7.57	13.6
Italy	5.61	3.77	7.71	13.2	45.1	8.98	15.6
France	4.47	3.01	5.19	10.6	59.6	6.34	10.8
Australia	6.19	4.16	8.1	17.3	39.6	9.19	15.5
Greece	5.5	3.69	6.38	13	49.4	8.79	13.3
Norway	5.38	3.61	6.24	13.3	47.5	8.6	15.3
South Korea	**10.1**	4.13	7.13	**20.6**	33	10.1	14.9
United States	**2.03**	1.36	**3.19**	10.5	72.8	**3.75**	**6.39**
Canada	5.64	3.79	8	14.3	45.8	8.81	13.7
Netherlands	5.62	3.78	**8.71**	15.2	41.7	8.98	16
Germany	4.9	3.6	6.64	12	54.1	6.95	11.9
United Kingdom	5.19	4.05	6.6	14.8	43.7	7.36	18.4
Russia	6.67	**4.48**	7.74	15.7	**27.3**	**10.5**	27.6
China	6.14	4.12	7.12	14.5	29	8.7	**30.4**
Target							
Port	4.5	2.68	**8.23**	13.5	51.2	6.53	13.3
Vessel	**7.72**	3.02	5.25	13.9	**28.5**	7.35	**34.2**
Other	**2.8**	**2.09**	**3.63**	**7.75**	**72.2**	**4.72**	**6.68**
Shipping company	2.9	2.62	4.24	9.77	66.5	6.65	7.36
Energy distributor	5.94	3.96	6.9	13.4	47.9	8.21	13.7
Shipbuilding firm	5.43	3.62	6.74	**17.5**	45	8.02	13.7
Offshore structure	6.35	**4**	6.59	13.5	47.3	**9.22**	13
Consequence							
Major	**3.97**	**2.62**	**3.27**	**11.8**	**66.4**	**5.74**	**6.19**
Minor	**5.61**	**3.39**	**8.86**	**12.4**	**38**	**8.42**	**23.4**

not only shapes the temporal distribution of threats but also significantly influences their spatial proliferation. Spoofing and Hacking threats also exhibit relatively high TRI values under the "Year" dimension: 23.865 and 21.38; their impact diminishes across geographical and target-related categories. This pattern suggests that their relevance is more temporally confined, rather than extending across broader spatial or functional dimensions. Additionally, DDoS and malware demonstrate more nuanced profiles. Both show moderately strong TRI scores in

Table 8
TRI of RIFs.

	Jamming	Phishing	DDoS	Hacking	Ransomware	Malware	Spoofing	Average
Year	11.835	8.605	14.36	21.38	37.81	10.775	23.865	18.376
Cyber threat origin	4.45	2.365	3.28	8.6	26.2	6.22	7.63	8.519
Region	2.74	0.96	1.575	3.015	12.75	1.32	5.07	3.727
Victim countries	4.035	1.65	2.76	7.265	23	3.375	12.005	7.014
Target	2.46	0.955	1.635	4.875	21.85	2.25	13.76	5.671
Consequence	0.82	0.385	2.795	0.3	14.2	1.34	8.605	3.307

temporal and origin dimensions, yet their influence across specific targets and consequences is considerably lower. This uneven distribution implies differentiated attack strategies and operational objectives within the broader cyber threat landscape.

Overall, the TRI value underscores the amplifying effect of temporal factors on the perceived risk influence of cyber threats, while spatial and operational dimensions serve to moderate these effects. A small cluster of RIFs, led by ransomware, exerts a disproportionate impact on the maritime cyber risk environment, highlighting the need for prioritised attention in both strategic planning and operational defence.

5. Discussion and implications

This section provides a comprehensive discussion and the implications of the proposed modelling framework through a series of comparative and scenario-based analyses. Section 5.1 provides a detailed comparative analysis of different types of cyber-attacks, systematically benchmarking the proposed approach against findings from previous studies. Section 5.2 presents scenario analyses that illustrate the model's applicability in real-world maritime cybersecurity contexts, focusing on ransomware incidents within the maritime sector and spoofing attacks targeting vessel systems. Finally, Section 5.3 discusses the broader implications of the findings, highlighting their relevance for risk assessment, decision support, and the development of resilient maritime cyber-defence strategies.

5.1. Comparative analysis on different types of cyber-attacks

Since this study reconstructs the BN model using an LLM approach, the proposed model represents a significant advancement over those used in prior research. Accordingly, this subsection conducts a comparative analysis between the updated model and previously reported findings related to various cyber-attack types, with the aim of highlighting the latest insights and methodological progress achieved in this study. By examining the updated BN model constructed from the LLM extraction dataset, it is evident that the ransomware threat remains the dominant cyber threat in the maritime domain. However, when the ransomware proportion is set to 100%, several distributional and dependency changes are observed compared to the previous model, reflecting evolving attack behaviours and geographical shifts, as shown in Fig. 5. In this scenario, ransomware primarily targets shipping companies (26.4%) and ports (14.6%). However, a notable increase in attacks against shipbuilding firms (8.8%) and offshore structures (8.2%) is also observed. This indicates that ransomware campaigns are no longer confined to operational logistics nodes but are expanding toward industrial and infrastructure sectors within the maritime ecosystem. From a temporal perspective, the likelihood of ransomware incidents peaks sharply in 2023 (36.7%) and 2022 (27.7%), suggesting an escalation of attacks in the post-pandemic period. The earlier concentration observed in 2020 has shifted toward more recent years, demonstrating the persistent and adaptive nature of ransomware actors. Regarding the geographical origin, Russia (22.7%) remains a major source, but there is a significant rise in incidents attributed to Unknown origins (61.1%),

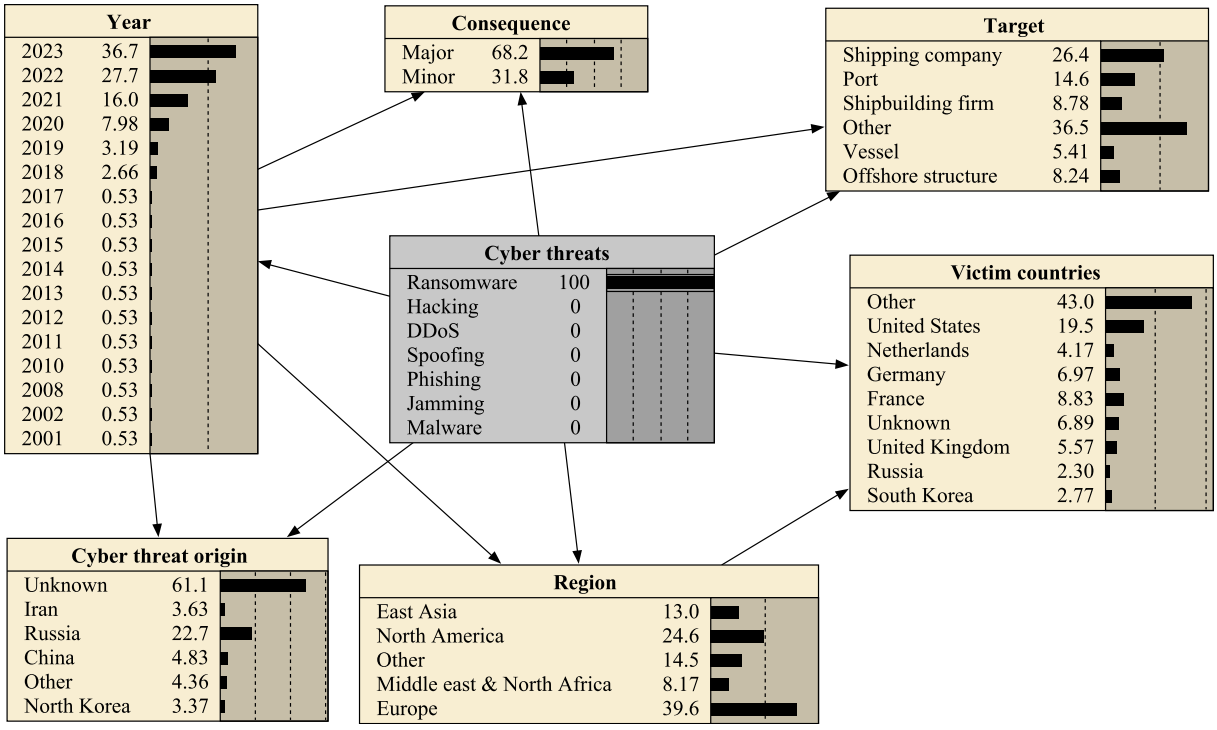

Fig. 5. Distribution of Ransomware attacks in the BN model from 2001 to 2023.

implying increasing use of anonymisation technologies or transnational cooperation among threat actors. Spatially, Europe (39.6%) continues to be the most affected region, followed by North America (24.6%), consistent with the previous trend. However, the presence of affected entities in East Asia (13.0%) and the Middle East & North Africa (8.2%) indicates a wider global exposure. On the victim-country level, the United States (19.5%) and Germany (6.97%) remain prominent targets, while the Netherlands, France, and South Korea newly emerge as affected nodes. In terms of consequences, about 68.2% of ransomware incidents lead to major outcomes, reflecting an even greater proportion of high-impact events compared with the prior model (60%). This underscores a growing operational and economic risk. Overall, the new model highlights a temporal shift toward 2022-2023, an expansion of target diversity, and an increasing difficulty in tracing attack origins, all of which suggest a trend toward more sophisticated and globally dispersed ransomware operations within the maritime sector.

On the other hand, when the DDoS attack type is set to 100%, as shown in Fig. 6, the updated BN model reveals notable temporal, spatial, and structural shifts in DDoS attack patterns within the maritime sector. From a temporal perspective, DDoS activity has intensified in recent years, with a marked surge in 2023 (27.3%) and 2022 (12.1%), replacing the earlier concentration around 2015. This pattern signifies a recent resurgence of DDoS campaigns, likely reflecting the growing exploitation of vulnerable maritime digital infrastructures and networked control systems. Regarding target preference, the attacks now focus more heavily on ports (29.4%) and shipping companies (17.0%), while shipbuilding firms (14.5%) and offshore structures (12.6%) also emerge as significant targets. This diversification indicates a shift from earlier operational systems toward broader maritime service and industrial infrastructures, implying higher systemic risk and cascading effects on logistics networks. Geographically, Europe (28.1%) and North America (23.8%) remain the most affected regions, consistent with the distribution of technologically advanced maritime hubs. The Middle East & North Africa (16.7%) and East Asia (14.6%) now show a more visible presence, suggesting an eastward extension of DDoS activities. On the victim-country level, the United States (15.7%), the Netherlands (11.7%), and Germany (9.97%) rank highest, indicating continued targeting of Western digital infrastructures, while Russia, South Korea, and France share similar exposure levels. In terms of attack origins, the share of incidents attributed to Unknown sources (25.1%) has risen markedly, whereas Russia (22.6%) remains a major contributor, followed by Iran (12.6%) and China (12.6%). This growing uncertainty in attribution reflects enhanced obfuscation techniques and possibly the use of global botnets for distributed attacks. Finally, the consequence distribution shows that minor incidents (65.3%) now outweigh major ones (34.7%), differing from ransomware trends. This aligns with DDoS's typical operational impact, temporary service disruption rather than permanent data loss, but the persistence of one-third major cases still signals significant operational and reputational risks.

In summary, while the earlier model depicted DDoS as a moderately distributed and somewhat localised threat, the updated model portrays it as a globally expanding, multifaceted, and increasingly anonymised phenomenon, with stronger emphasis on ports, diversified regional impact, and sustained operational consequences in the maritime domain.

5.2. Scenario analysis

5.2.1. Ransomware attack on the maritime sector

When "Cyber threat origin" was set to Unknown, "Region" to Europe, "Consequence" to Major, and "Victim countries" to United Kingdom (all at 100%), the probability of Ransomware attacks sharply increased from 54.3% in the baseline model to 84.5%, while Hacking, Spoofing, and Malware declined to 8.2%, 2.93%, and 1.11%, respectively. As shown in Fig. 7. This overwhelming rise identifies ransomware as the primary cyber threat responsible for severe impacts in the UK maritime sector. Such a pattern underscores the growing dominance of extortion-driven attacks that paralyse port operations, disrupt digital logistics systems, and cause large-scale economic losses. The "Unknown" origin likely reflects sophisticated threat actors employing obfuscation techniques, such as proxy servers, encryption, and cryptocurrency payments, to conceal their identities. In this context, the findings suggest that ransomware campaigns against UK maritime assets are strategically designed to maximise operational disruption while minimising

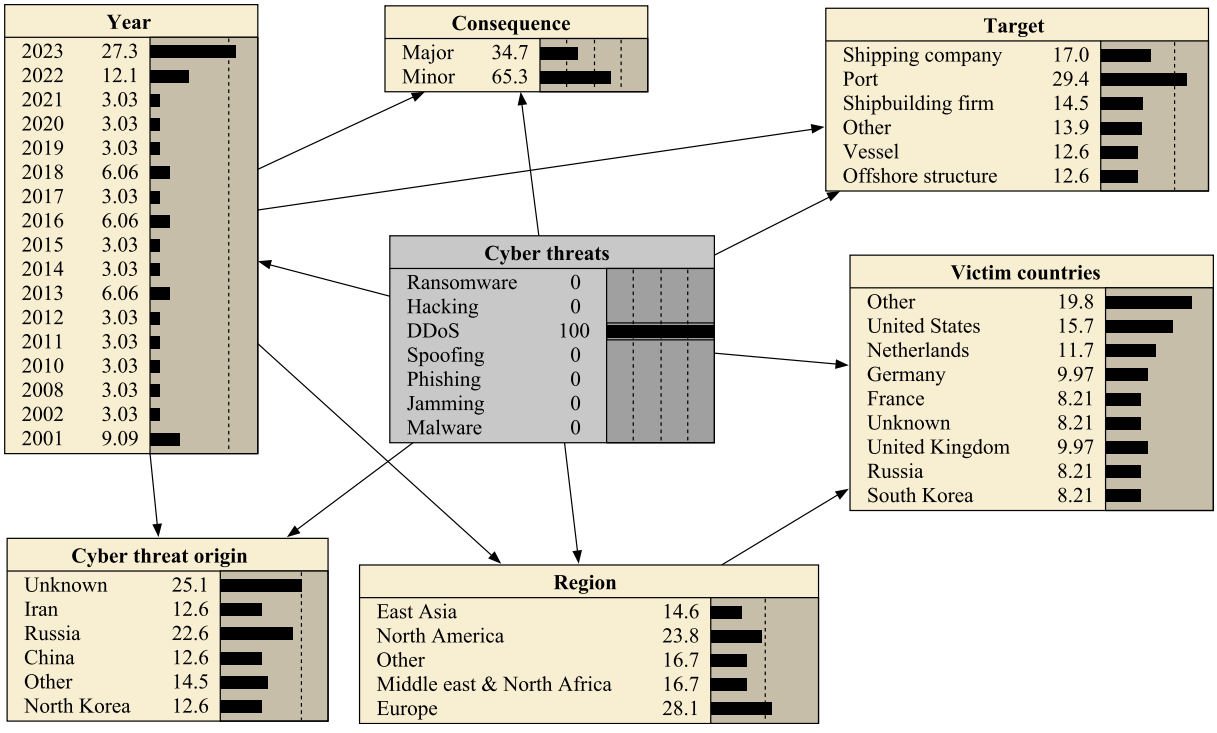

Fig. 6. Distribution of DDoS attacks in the BN model from 2001 to 2023.

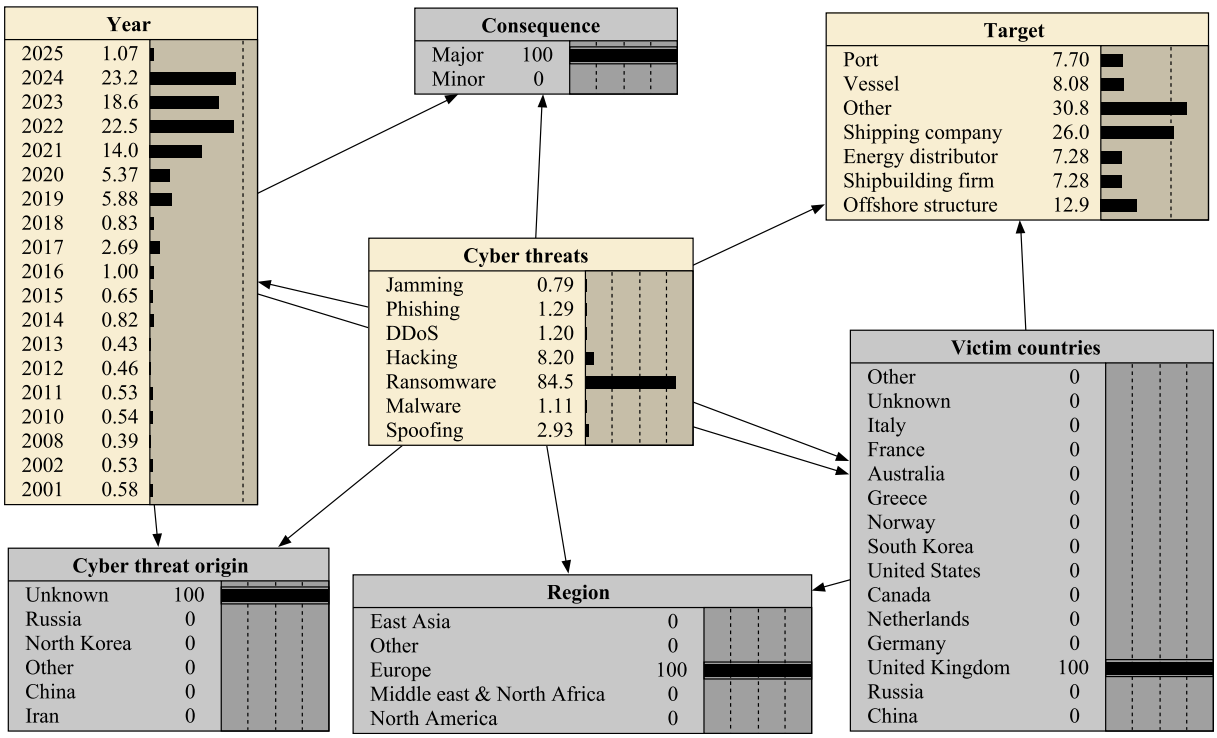

Fig. 7. Scenario one: Ransomware Attack on the Maritime Sector.

traceability.

Under the same conditions, the Target node exhibited a major redistribution. The likelihood of attacks against Other onshore maritime targets rose from 20.6% to 30.8%, and against Shipping companies from 20.0% to 26.0%, while Vessels and Offshore structures decreased to 8.08% and 12.9%, respectively. This transition indicates that ransomware attacks of unknown origin focus predominantly on digitalised and onshore infrastructures rather than on vessels or offshore maritime assets. These may include enterprise management systems, logistics scheduling platforms, or port information networks that connect multiple maritime stakeholders. Such targets are typically more exposed through corporate networks and third-party service dependencies. The decline in vessel-related attacks suggests that onboard systems may be better isolated or protected by stricter security controls, whereas the

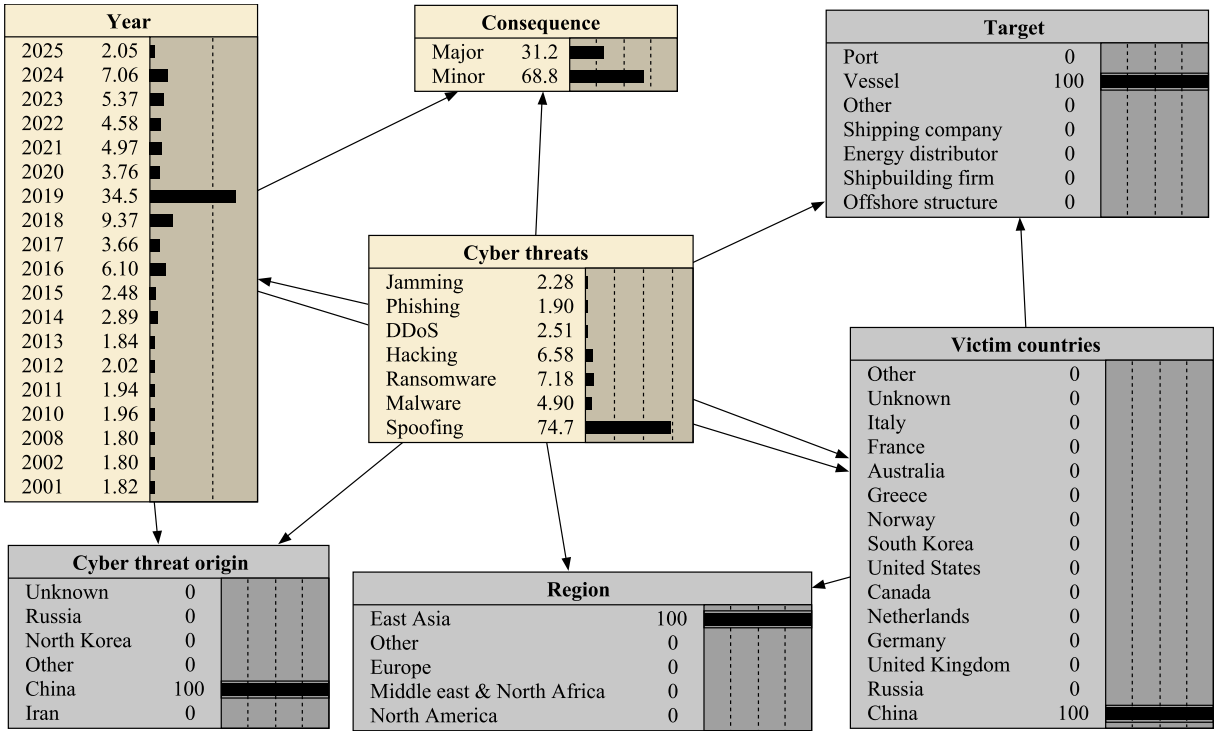

Fig. 8. Scenario two: Spoofing Attack on Vessel Systems.

interconnected back-end systems of shipping firms remain more vulnerable to exploitation.

The Year node shows a significant shift toward 2021 (14%) and 2022 (22.5%), whereas previous years exhibit much lower probabilities. Compared with the baseline distribution, this indicates that ransomware incidents of unknown origin targeting the UK maritime sector have surged in recent years. This temporal clustering mirrors the global escalation of ransomware attacks since 2020, coinciding with the rapid digital transformation of shipping and port operations. The trend may also be linked to growing geopolitical tensions, where anonymous actors exploit cyber vulnerabilities under the guise of financial motivation. From a predictive standpoint, this concentration suggests that future maritime cybersecurity strategies must prioritise adaptive defence mechanisms, real-time anomaly detection, and cross-sector intelligence sharing to mitigate such rapidly evolving threats.

5.2.2. Spoofing attack on vessel systems

When "Cyber threat origin" was set to China, "Region" to East Asia, "Target" to Vessel, and "Victim countries" to China (all at 100%), the probability of Spoofing attacks rose dramatically from 13.5% in the baseline model to 74.7%, surpassing all other attack categories. As shown in Fig. 8. This steep increase indicates that navigation and positioning system manipulation is the most probable cyber threat in this context. Spoofing attacks typically target navigation and positioning systems such as GPS or AIS (Automatic Identification System). They can lead to navigation errors, false route displays, or tracking disruptions, and in some cases, attackers may deliberately disable GPS systems to achieve economic or strategic objectives, such as concealing real routes, evading monitoring, or creating confusion in congested waters. The significant rise in spoofing probability reflects both the operational exposure of vessels to satellite-based navigation interference and the technical feasibility of such attacks in Chinese coastal and inland waters. It also implies that the maritime sector in China faces persistent risks from signal manipulation rather than direct data encryption or ransom-driven threats.

The Consequence node shows a notable shift from Major (57.4%) to 31.2%, and Minor (42.6%) to 68.8%. This indicates that most spoofing-related incidents in this scenario result in light or short-term operational impacts, rather than severe consequences. A higher proportion of Minor outcomes suggests that these attacks often cause temporary disruptions, such as false location readings or short-term system unavailability, without leading to financial loss or equipment damage. This pattern also aligns with the nature of spoofing, where the immediate effects are operational and reversible, unless combined with secondary attacks. It highlights that Chinese vessel cyber incidents are frequent but manageable, emphasising the need for real-time detection and rapid correction rather than large-scale recovery operations.

The Year node exhibits a sharp increase for 2019, rising from 5.38% in the baseline to 34.5% after conditioning on Chinese vessel scenarios. This suggests that spoofing-related cyber events were most likely to occur around 2019. The spike in 2019 may correspond to a period of rapid maritime digitalisation in China, during which vessel navigation systems became more interconnected with port and satellite communication infrastructures. The integration of new technologies such as BeiDou positioning systems and smart port logistics platforms may have increased exposure to cyber vulnerabilities, especially in the transition from legacy systems to modern digital architectures.

5.3. Implications

The scenarios detailed above highlight the complex nature of cyber threats in the maritime domain, with important implications for cybersecurity strategy, operational resilience, and risk governance. Rather than representing isolated technical incidents, the analysed cyber events demonstrate how cyberattacks can disrupt maritime operations and affect interconnected socio-technical systems. The findings also suggest that cyberattacks may be associated with broader economic and geopolitical contexts beyond purely technical motivations. These findings provide several practical implications for maritime cybersecurity governance and risk management.

First, the prevalence of attack types such as hacking, spoofing, and ransomware across different geopolitical contexts underscores the need to broaden the maritime sector's conceptualisation of cyber threats. Traditional frameworks have primarily focused on technical vulnerabilities and system-level faults. However, as demonstrated by the 2016 and 2022 scenarios, cyber threats increasingly operate at the intersection of political intent, strategic deterrence, and technological exploitation, with direct implications for crisis escalation and emergency response effectiveness. This necessitates a paradigm shift from purely system-centric protection towards integrated risk assessment approaches that explicitly account for geopolitical, socio-technical, and economic drivers of cyber-induced emergencies.

Second, the critical targeting of ports and vessels across these scenarios highlights the vulnerability of maritime infrastructure as soft targets in cyber conflicts and potential focal points of cascading emergencies. Ports, serving as linchpins in global trade networks, present single points of failure whose disruption can rapidly propagate across supply chains, emergency response capabilities, and national economies. For countries such as China and the United States, whose economic resilience is deeply intertwined with maritime logistics, these vulnerabilities translate directly into system-level crisis risks rather than isolated cybersecurity concerns. Accordingly, cybersecurity resilience must be treated as an integral component of national and sectoral emergency management strategies, particularly for assets designated as critical infrastructure.

Third, the persistent challenge of attribution, particularly evident in the 2016 scenario where the threat origin was classified as Unknown, undermines effective incident response, crisis communication, and strategic deterrence. In emergency contexts, attribution ambiguity delays coordinated response actions, complicates cross-border information sharing, and weakens accountability mechanisms. This opacity benefits malicious actors by enabling plausible deniability and prolonging uncertainty during crisis situations. Addressing this challenge requires enhanced international cooperation on cyber attribution protocols, shared threat intelligence infrastructures, and harmonised legal frameworks to support coordinated crisis response and recovery. Maritime nations must therefore pursue diplomatic efforts to establish cyber norms while simultaneously strengthening technical and organisational capacity for forensic attribution.

Fourth, the escalation of ransomware attacks in the 2022 Russia-origin scenario reflects a broader transition towards economically motivated cyber operations that can precipitate large-scale operational crises. These attacks impose both direct financial losses and indirect systemic impacts, including reputational damage, operational paralysis, and erosion of trust in digital maritime services. From a resilience perspective, such attacks expose the fragility of digitally dependent emergency response and logistics coordination mechanisms. Addressing these risks requires not only enhanced preventive controls, such as endpoint protection and network segmentation, but also system-level resilience measures, including cyber insurance mechanisms, redundancy planning, and resilience-by-design principles embedded across maritime supply chains.

Finally, the observed divergence in threat types across countries and time periods underscores that maritime cyber risk is inherently context-dependent, shaped by geopolitical conditions, technological maturity, and organisational preparedness. Risk assessment frameworks based on limited historical evidence may provide only a partial representation of maritime cybersecurity risks, particularly during cyber incidents that require timely situational awareness and effective decision-making. These findings highlight the importance of scalable and extensible risk assessment approaches that can systematically incorporate newly available incident information and threat patterns. The application of

LLM-enabled analytical frameworks offers significant potential to improve the efficiency and consistency of cyber-risk information extraction and support data-driven risk assessment in complex maritime environments.

In summary, these scenarios reveal a maritime cybersecurity landscape increasingly characterised by cyber-induced crises, asymmetric threat strategies, and cascading system-level consequences. Addressing these challenges requires more than incremental technical improvements, it demands a reconfiguration of maritime cyber risk governance that explicitly links cybersecurity with crisis preparedness, emergency response, and resilience management. Future research and policy development should therefore prioritise integrated frameworks that bridge technical, organisational, and geopolitical dimensions, enabling context-sensitive, decision-relevant, and resilience-oriented approaches to managing cyber risks in complex maritime systems.

6. Conclusions

This research examined the integration of LLM-based information extraction and BN modelling for maritime cybersecurity risk assessment. The findings indicate that information derived from incident reports can be transformed into structured inputs for probabilistic risk analysis, reducing reliance on manual classification and expert interpretation. The resulting models capture dependencies among cyber-risk factors and support the identification of key relationships associated with maritime cyber incidents.

From a practical perspective, the framework reduces reliance on labour-intensive manual coding while supporting systematic analysis of cyber-risk relationships and propagation mechanisms. The resulting BN provides an interpretable decision-support tool for evaluating cyber-risk scenarios, identifying influential risk factors, and supporting evidence-based cybersecurity management in the maritime sector. More broadly, the framework offers a transferable methodology for safety–critical domains where risk information is primarily embedded within unstructured accident or incident reports.

Several limitations should be acknowledged. First, the dependency structure learned by the TAN algorithm reflects statistical associations derived from observational data and should not be interpreted as fully validated causal relationships. Second, the framework was evaluated using a historical dataset of publicly reported incidents, which may be affected by reporting bias, underreporting, and variations in disclosure practices. Third, although the LLM-based extraction process demonstrated high accuracy and consistency, its performance remains dependent on the quality and completeness of the underlying incident narratives.

Future research should investigate the integration of additional data sources, including operational cybersecurity logs and threat intelligence feeds, to improve data completeness and model robustness. Further work may also explore online Bayesian learning, dynamic risk assessment, and domain-adapted or open-source LLMs to enhance transparency, adaptability, and real-time decision-support capabilities.

This study demonstrates the potential of integrating LLMs with probabilistic risk modelling to support data-driven and interpretable cybersecurity risk assessment. The proposed framework contributes to the growing application of AI in safety and risk analysis and provides a foundation for advancing cyber-risk governance, resilience assessment, and emergency management in increasingly digitalised socio-technical systems.

CRediT authorship contribution statement

Yunfeng Zhao: Writing – review & editing, Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Huanhuan Li:** Writing – review & editing, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Trung Thanh Nguyen:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Christian Matthews:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Zaili Yang:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Zaili Yang reports financial support was provided by European Research Council.

Acknowledgement

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 864724) and by LJMU Freeport Thematic Doctoral Pathways programme.

Data availability

Data will be made available on request.

References

Ahmadi, E., Muley, S., Wang, C., 2025. Automatic construction accident report analysis using large language models (LLMs). J. Intelligent Construction 3, 1–10. https://doi.org/10.26599/JIC.2024.9180039.

Androjna, A., Brcko, T., Pavic, I., Greidanus, H., 2020. Assessing cyber challenges of maritime navigation. J. Marine Sci. Eng. 8, 776. https://doi.org/10.3390/jmse8100776.

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

Badea, M., Bucovețchi, O., Gheorghe, A.V., Hnatiuc, M., Raicu, G., 2025. Maritime industry cybersecurity threats in 2025: advanced persistent threats (APTs), Hacktivism and Vulnerabilities. Logistics 9. https://doi.org/10.3390/logistics9040178.

Baybutt, P., 2015. A critique of the Hazard and Operability (HAZOP) study. J. Loss Prev. Process Ind. 33, 52–58. https://doi.org/10.1016/j.jlp.2014.11.010.

Bayer, M., Kuehn, P., Shanehsaz, R., Reuter, C., 2024. CySecBERT: a domain-adapted language model for the cybersecurity domain. ACM Trans. Priv. Secur. 27, 20. https://doi.org/10.1145/3652594.

Ben Farah, M.A., Ukwandu, E., Hindy, H., Brosset, D., Bures, M., Andonovic, I., Bellekens, X., 2022. Cyber security in the maritime industry: a systematic survey of recent advances and future trends. Information 13, 22. https://doi.org/10.3390/info13010022.

Benmalek, M., 2024. Ransomware on cyber-physical systems: taxonomies, case studies, security gaps, and open challenges. Internet Things Cyber-Phys. Syst. 4, 186–202. https://doi.org/10.1016/j.iotcps.2023.12.001.

Bolbot, V., Theotokatos, G., Boulougouris, E., Vassalos, D., 2020. A novel cyber-risk assessment method for ship systems. Saf. Sci. 131, 104908. https://doi.org/10.1016/j.ssci.2020.104908.

Chaal, M., Ren, X., BahooToroody, A., Basnet, S., Bolbot, V., Banda, O.A.V., Gelder, P.V., 2023. Research on risk, safety, and reliability of autonomous ships: a bibliometric review. Saf. Sci. 167, 106256. https://doi.org/10.1016/j.ssci.2023.106256.

Chang, C.-H., Kontovas, C., Yu, Q., Yang, Z., 2021. Risk assessment of the operations of maritime autonomous surface ships. Reliab. Eng. Syst. Saf. 207, 107324. https://doi.org/10.1016/j.ress.2020.107324.

Cheung, K.-F., Bell, M.G.H., Bhattacharjya, J., 2021. Cybersecurity in logistics and supply chain management: an overview and future research directions. Transp. Res. Part E: Logistics and Transp. Rev. 146, 102217. https://doi.org/10.1016/j.tre.2020.102217.

Clavijo Mesa, M.V., Patino-Rodriguez, C.E., Guevara Carazas, F.J., 2024. Cybersecurity at sea: a literature review of cyber-attack impacts and defenses in maritime supply chains. Information 15, 710. https://doi.org/10.3390/info15110710.

Deng, Z., Ma, W., Han, Q.-L., Zhou, W., Zhu, X., Wen, S., Xiang, Y., 2025. Exploring DeepSeek: a survey on advances, applications, challenges and future directions. IEEE/CAA J. Autom. Sin. 12, 872–893. https://doi.org/10.1109/JAS.2025.125498.

Du, G., Chen, A., 2025. Coal mine accident risk analysis with large language models and bayesian networks. Sustainability 17, 1896. https://doi.org/10.3390/su17051896.

Ferdous, R., Khan, F., Sadiq, R., Amyotte, P., Veitch, B., 2011. Fault and event tree analyses for process systems risk analysis: uncertainty handling formulations. Risk Anal. 31, 86–107. https://doi.org/10.1111/j.1539-6924.2010.01475.x.

Gholamy, A., Kreinovich, V., Kosheleva, O., 2018. Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation. Int. J. Intell. Technol. Appl. Stat. 11, 105–111. https://doi.org/10.6148/IJITAS.201806_11(2).0003.

Guidelines on cyber security onboard ships [WWW Document], 2024. URL https://www.bimco.org/news-insights/bimco-news/2024/20241114-cyber-security-guidelines/ (accessed 12.2.25).

Hanafiah, R.M., Abdullah, M.A., Zaideen, I.M.M., Najib, A.F.A., Rahman, N.S.F.A., Karim, N.H., 2025. Selection of the regulatory seaport cybersecurity based on integrated Ahp and Topsis. Asian Academy of Manage. J. 20, 1–33. https://doi.org/10.21315/aamj2025.30.l.1.

Harish, A.V., Tam, K., Jones, K., 2025. Literature review of maritime cyber security: the first decade. Maritime Technol. Res. 7, 273805. https://doi.org/10.33175/mtr.2025.273805.

Hasanov, I., Virtanen, S., Hakkala, A., Isoaho, J., 2024. Application of large language models in cybersecurity: a systematic literature review. IEEE Access 12, 176751–176778. https://doi.org/10.1109/ACCESS.2024.3505983.

Hassanin, M., Keshk, M., Salim, S., Alsubaie, M., Sharma, D., 2025. PLLM-CS: pre-trained Large Language Model (LLM) for cyber threat detection in satellite networks. Ad Hoc Netw. 166, 103645. https://doi.org/10.1016/j.adhoc.2024.103645.

Henriques de Gusmão, A.P., Mendonça Silva, M., Poleto, T., Camara e Silva, L., Cabral Seixas Costa, A.P., 2018. Cybersecurity risk analysis model using fault tree analysis and fuzzy decision theory. Int. J. Inf. Manag. 43, 248–260. https://doi.org/10.1016/j.ijinfomgt.2018.08.008.

Huang, L., Xiao, X., 2024. CTIKG: LLM-powered knowledge graph construction from cyber threat intelligence. In: First Conference on Language Modeling.

Li, H., Zhou, K., Zhang, C., Bashir, M., Yang, Z., 2024. Dynamic evolution of maritime accidents: comparative analysis through data-driven Bayesian networks. Ocean Eng. 303, 117736. https://doi.org/10.1016/j.oceaneng.2024.117736.

Li, K.X., Li, M., Zhu, Y., Yuen, K.F., Tong, H., Zhou, H., 2023. Smart port: a bibliometric review and future research directions. Transp. Res. Part E: Logistics and Transp. Rev. 174, 103098. https://doi.org/10.1016/j.tre.2023.103098.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G., 2023. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. 55, 35. https://doi.org/10.1145/3560815.

Liu, X., Hu, J., Mei, Q., Wang, S., 2026. Pirate-GPT: a locally deployed large language model framework for reliable offline anti-piracy decision support and knowledge retrieval in maritime operations. Reliab. Eng. Syst. Saf. 267, 111891. https://doi.org/10.1016/j.ress.2025.111891.

Mao, R., Chen, G., Zhang, X., Guerin, F., Cambria, E., 2024. GPTEval: A survey on assessments of ChatGPT and GPT-4. Doi:10.48550/arXiv.2308.12488.

Miller, T., Durlik, I., Kostecka, E., Sokołowska, S., Kozlovska, P., Zwolak, R., 2025. Artificial intelligence in maritime cybersecurity: a systematic review of AI-driven threat detection and risk mitigation strategies. Electronics 14. https://doi.org/10.3390/electronics14091844.

Mohsendokht, M., Li, H., Kontovas, C., Chang, C.-H., Qu, Z., Yang, Z., 2026. Systemic risk analysis of complex socio-technical systems from the safety-II perspective. Reliab. Eng. Syst. Saf. 270, 112200. https://doi.org/10.1016/j.ress.2026.112200.

Mohsendokht, M., Li, H., Kontovas, C., Chang, C.-H., Qu, Z., Yang, Z., 2024. Decoding dependencies among the risk factors influencing maritime cybersecurity: lessons learned from historical incidents in the past two decades. Ocean Eng. 312, 119078. https://doi.org/10.1016/j.oceaneng.2024.119078.

Motlagh, F.N., Hajizadeh, M., Majd, M., Najafi, P., Cheng, F., Meinel, C., 2024. Large language models in cybersecurity: state-of-the-art. Doi:10.48550/arXiv.2402.00891.

Mudassar Yamin, M., Hashmi, E., Ullah, M., Katt, B., 2024. Applications of LLMs for generating cyber security exercise scenarios. IEEE Access 12, 143806–143822. https://doi.org/10.1109/ACCESS.2024.3468914.

Nguyen, S., Shu-Ling Chen, P., Du, Y., 2022. Risk assessment of maritime container shipping blockchain-integrated systems: an analysis of multi-event scenarios. Transp. Res. Part E: Logistics and Transp. Rev. 163, 102764. https://doi.org/10.1016/j.tre.2022.102764.

Pang, Z., Luan, Y., Chen, J., Li, T., 2024. ParInfoGPT: an LLM-based two-stage framework for reliability assessment of rotating machine under partial information. Reliab. Eng. Syst. Saf. 250, 110312. https://doi.org/10.1016/j.ress.2024.110312.

Park, C., Kontovas, C., Yang, Z., Chang, C.-H., 2023. A BN driven FMEA approach to assess maritime cybersecurity risks. Ocean & Coastal Manage. 235, 106480. https://doi.org/10.1016/j.ocecoaman.2023.106480.

Peng, P., Xie, X., Claramunt, C., Lu, F., Gong, F., Yan, R., 2025. Bibliometric analysis of maritime cybersecurity: research status, focus, and perspectives. Transp. Res. Part E: Logistics and Transp. Rev. 195, 103971. https://doi.org/10.1016/j.tre.2025.103971.

Pham, C.M., Hoyle, A., Sun, S., Resnik, P., Iyyer, M., 2024. TopicGPT: A prompt-based topic modeling framework. Doi:10.48550/arXiv.2311.01449.

Pijpker, J., McCombie, S., Johnson, S., Loves, R., Makrakis, G.M., 2024. An open-source database of cyberattacks on the maritime transportation system.

Potamos, G., Stavrou, E., Stavrou, S., 2024. Enhancing maritime cybersecurity through operational technology sensor data fusion: a comprehensive survey and analysis. Sensors 24, 3458. https://doi.org/10.3390/s24113458.

Powers, D., Ailab, 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. 2, 2229–3981. https://doi.org/10.9735/2229-3981.

Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A., 2025. A systematic survey of prompt engineering in large language models: techniques and applications. Doi:10.48550/arXiv.2402.07927.

Schauer, S., Polemi, N., Mouratidis, H., 2019. MITIGATE: a dynamic supply chain cyber risk assessment methodology. J. Transp. Secur. 12, 1–35. https://doi.org/10.1007/s12198-018-0195-z.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.

Tam, K., Jones, K., 2019. MaCRA: a model-based framework for maritime cyber-risk assessment. WMU J. Marit. Aff. 18, 129–163.

Tam, K., Jones, K., 2018. Cyber-Risk Assessment for Autonomous Ships, in: 2018 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). Presented at the 2018 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), pp. 1–8. Doi:10.1109/CyberSecPODS.2018.8560690.

Tusher, H.M., Munim, Z.H., Notteboom, T.E., Kim, T.-E., Nazir, S., 2022. Cyber security risk assessment in autonomous shipping. Marit Econ Logist 24, 208–227. https://doi.org/10.1057/s41278-022-00214-0.

Uflaz, E., Sezer, S.I., Tunçel, A.L., Aydin, M., Akyuz, E., Arslan, O., 2024. Quantifying potential cyber-attack risks in maritime transportation under Dempster–Shafer theory FMECA and rule-based Bayesian network modelling. Reliab. Eng. Syst. Saf. 243, 109825. https://doi.org/10.1016/j.ress.2023.109825.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., Wang, G., 2023. GPT-NER: Named entity recognition via large language models. Doi:10.48550/arXiv.2304.10428.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D., 2023. Self-consistency improves chain of thought reasoning in language models. Doi:10.48550/arXiv.2203.11171.

Wang, Z., Pang, Y., Lin, Y., 2023. Large language models are zero-shot text classifiers. Doi:10.48550/arXiv.2312.01044.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q.V., Zhou, D., 2022. Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Proces. Syst. 35, 24824–24837.

Wei, M., Cui, Y., Liu, J., 2026. Unveiling the influencing factors of maritime accidents through data-driven approaches: leveraging large language model tools. Saf. Sci. 197, 107116. https://doi.org/10.1016/j.ssci.2026.107116.

Xu, H., Wang, S., Li, N., Wang, K., Zhao, Y., Chen, K., Yu, T., Liu, Y., Wang, H., 2025. Large language models for cyber security: a systematic literature review. ACM Trans. Softw. Eng. Methodol. https://doi.org/10.1145/3769676.

Yang, Z., Bonsall, S., Wang, J., 2008. Fuzzy rule-based bayesian reasoning approach for prioritization of failures in FMEA. IEEE Trans. Reliab. 57, 517–528. https://doi.org/10.1109/TR.2008.928208.

Yoo, Y., Park, H.-S., Yoo, Y., Park, H.-S., 2021. qualitative risk assessment of cybersecurity and development of vulnerability enhancement plans in consideration of digitalized ship. J. Marine Sci. Eng. 9. https://doi.org/10.3390/jmse9060565.

Yu, H., Meng, Q., Fang, Z., Liu, J., 2023. Literature review on maritime cybersecurity: state-of-the-art. J. Navigation 76, 453–466. https://doi.org/10.1017/S0373463323000164.

Zagan, R., Raicu, G., Sabau, A., 2022. Studies and research regarding vulnerabilities of marine autonomous surface systems (mass) and remotely operated vessels (rovs) from point of view of cybersecurity. Int. J. Modern Manuf. Technol. 14, 310–318.

Zhang, D., Yan, X., Zhang, J., Yang, Z., Wang, J., 2016. Use of fuzzy rule-based evidential reasoning approach in the navigational risk assessment of inland waterway transportation systems. Saf. Sci. 82, 352–360. https://doi.org/10.1016/j.ssci.2015.10.004.

Zhang, D., Yan, X.P., Yang, Z.L., Wall, A., Wang, J., 2013. Incorporation of formal safety assessment and Bayesian network in navigational risk estimation of the Yangtze River. Reliab. Eng. Syst. Saf. 118, 93–105. https://doi.org/10.1016/j.ress.2013.04.006.

Zhang, H., Wu, T., Zhu, T., Wen, S., Xiang, Y., 2025. CyberLLaMA: a fine-tuned large language model for cybersecurity named entity recognition. Knowl.-Based Syst. 328, 114183. https://doi.org/10.1016/j.knosys.2025.114183.

Zhang, Y., Liu, J., Zhong, X., Wu, L., 2025. SecLMNER: a framework for enhanced named entity recognition in multi-source cybersecurity data using large language models. Expert Syst. Appl. 271, 126651. https://doi.org/10.1016/j.eswa.2025.126651.

Zhen, H., Yang, J.J., 2025. CrashSage: A Large Language Model-Centered Framework for Contextual and Interpretable Traffic Crash Analysis. Doi:10.48550/arXiv.2505.07853.